Perceptual Flexibility in Native and Non-native Speech Perception and Sentence Processing: Listener's Attention Shifting across Speech Units and Attention Weight on Cues

Xiaomu Ren

BA, MA

SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR

THE DEGREE OF

DOCTOR OF PHILOSOPHY

SCHOOL OF CRITICAL STUDIES

COLLEGE OF ARTS

SUBMITTED SEPTEMBER 2024

REVISED FEBRUARY 2025

Abstract

Two overarching questions within the realm of speech perception and comprehension have been asked: Why is perceiving native speech effortless, while processing non-native speech is challenging? In this thesis, I address these questions by exploring how native and non-native listeners differ in their perceptual flexibility. I apply the concept of *perceptual flexibility* as a cognitive mechanism of listeners to shed light on the potential differences demonstrated by native and non-native listeners. This mechanism refers to how listeners adapt to variations in acoustic signals during speech perception and comprehension. Listeners with high perceptual flexibility are predicted to actively shift attention across and detect speech units of different sizes during real-time speech unit perception, and to assign attention weights to and flexibly integrate multiple cues during real-time sentence processing. In contrast, listeners with reduced perceptual flexibility are predicted to be less able to actively shift attention across units of different sizes and to rely consistently on certain cues with a reduced ability to integrate multiple cues. Therefore, I pose a general question: Do non-native listeners exhibit comparable perceptual flexibility to native listeners in speech unit perception and sentence processing? This thesis employs monitoring and visual-world eye-tracking paradigms to test listeners' perceptual flexibility, which is manifested in both low-level speech perception—detecting perceptual units of different sizes—and high-level sentence processing—utilizing and integrating multiple cues.

Experiments 1 and 2 are phoneme and syllable monitoring experiments using English (Experiment 1) and Mandarin (Experiment 2) pseudo-words. In each experiment there was a native listener group (English native speakers in Experiment 1, Mandarin native speakers in Experiment 2) and a non-native group (Mandarin native speakers learning English as L2 in Experiment 1, English native speakers learning Mandarin as L2 in Experiment 2). To observe how listeners flexibly shift their attention between phonemes and syllables, I manipulated the stimuli in two ways. First, I created stimuli with an artificial accent,

which served as bottom-up information. Second, I used prior knowledge of the artificial accent as top-down information. I then examined how reaction times (RTs) to phoneme and syllable targets changed under these manipulations. The results revealed that both English and Mandarin listeners exhibited comparable perceptual flexibility in detecting units of different sizes in the English context (Experiment 1), with greater sensitivity to syllables than to phonemes. In the Mandarin context (Experiment 2), English listeners showed higher perceptual flexibility than Mandarin listeners. These findings shed light on how perceptual units are represented and retrieved during native and non-native speech unit perception, providing experimental support that syllable processing is more easily disrupted than phoneme processing, especially by distorted bottom-up input, suggesting that phonemes may be a more robust unit than syllables during speech perception.

While Experiments 1 and 2 examined perceptual flexibility in perceiving sublexical units, Experiments 3 and 4 looked at sentence-level processing. Experiments 3 and 4 employed a visual-world eye-tracking paradigm with English (Experiment 3) and Mandarin (Experiment 4) sentence stimuli, and utilized native and non-native listener groups as for Experiments 1 and 2. Experiments 3 and 4 investigated how native and non-native listeners adopt different perceptual strategies during real-time sentence processing, focusing on how they allocate attention to various cues and integrate them flexibly. This eye-tracking experiment particularly focused on prosodic cues, verb semantics, and whether information is repeated or new in a broader discourse context. Results showed a complex three-way interaction among L1, prosody, and verb semantics. With both English and Mandarin stimuli, when processing sentences containing old information, non-native listeners tended to adopt perceptual strategies similar to those of native listeners. Specifically, both native and non-native listeners in both languages exhibited effect of both semantics and prosody, with the semantic effect being exaggerated when the target carried a prosodic accent. However, when processing sentences containing new information, non-native and native listeners exhibited divergent patterns, the specifics of which depended on whether the stimulus language was English or Mandarin. Native listeners can interactively combine both semantic and prosodic information across a variety of contexts. Non-native listeners can also do this, but only with familiar repeated information. In contexts when

they must handle new information, their ability to integrate different cues is reduced.

Thus, the main findings of the thesis are threefold. First, in the perception of low-level sublexical speech units, native and non-native listeners generally demonstrated comparable levels of perceptual flexibility when shifting their focus flexibly and actively between phonemes and syllables, with syllable processing being more susceptible to disruption than phoneme processing. Second, in high-level sentence processing, native and non-native listeners began to show a divergence in their ability to utilize and integrate multiple cues, particularly when processing sentences containing new information. Third, during high-level sentence processing, perceptual flexibility was influenced not only by a listener's linguistic experience but also by the characteristics of the language they were processing.

Acknowledgement

This PhD journey has been amazing, allowing me to experience a wonderful camaraderie with the people I worked with at Glasgow. First and foremost, I would like to express my eternal gratitude to my inspiring supervisory team of Dr Clara Cohen and Dr Rachel Smith, who supported me throughout this endeavour. Clara and Rachel are exceptionally professional and also patient in guiding me to complete and polish this project. They taught me the intricate experiment design and elevated my proficiency in high-level statistical analysis, and introduced me to various professional opportunities. Their professional expertise and patience have left an indelible mark on my academic growth, for which I am profoundly grateful.

I would also like to express special thanks to my parents, who have consistently supported me with their abundant love, both financially and emotionally. This work is also dedicated to the memory of my grandparents—Daoming Zhang and Chengfu Ren. I am eternally grateful for their unwavering support and guidance throughout my educational journey and for their attentive care in every aspect of my life.

Declaration

I declare that, except where explicit reference is made to the contribution of others, that this thesis is the result of my own work and has not been submitted for any other degree at the University of Glasgow or any other institution.

Xiaomu Ren

To my parents, Yanxia Du and Jun Ren

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Perceptual flexibility as a concept in speech perception

Outside the scope of speech perception, flexibility of human perception and cognition generally describes people's adaptable control over their perceptual experiences (Blake & Palmisano, 2021). The *rabbit-duck illusion* can be a typical example to illustrate the flexibility of the visual system (Blake & Palmisano, 2021; Freeman et al., 2022).

The rabbit-duck illusion is an ambiguous image that can be perceived as either a rabbit or a duck. The flexible nature of perception is evident as individuals can shift their interpretation of the image between a rabbit and a duck, demonstrating that our minds possess the ability to control and alter how we perceive stimuli. This phenomenon highlights the subjectivity of perception, as people do not always perceive stimuli in the same way and have flexible control over their perception, showcasing that human perceptions are not rigid or fixed but can be actively influenced and adjusted based on contextual factors and individual perspectives.

In linguistic contexts, the term *flexibility* was used in previous studies to describe how it supports and facilitates listeners to discriminate ambiguous sounds; to understand foreign accents, and acquire new languages. For example, to discriminate ambiguous sounds, the stability in the long-term language representation of speech is required to be flexible to accommodate talker-specific variations. The phonemic boundaries are demonstrated to

Figure 1.1: Duck-rabbit illusion. Reprinted from McManus et. al (2010)

be flexible and can adapt to variation in the signal (Kraljic & Samuel, 2005). Therefore, this cognitive flexibility in perception plays an important role in sentence processing, allowing listeners to tune in the ways talkers speak and to understand them better (Bent et al., 2016). For instance, Norris et al (2003) conducted experiments wherein one group of Dutch listeners heard an ambiguous sound between /f/ and /s/ in Dutch words ending in /f/, while another group heard the same sound in words ending in /s/. The results indicated that the first group learned to categorize the ambiguous sounds as /f/ rather than /s/, and conversely, the second group tended to categorize them as /s/ rather than /f/. This indicated that native listeners were able to mold their phonemic categories to reflect the phonetic elements they were currently confronted with.

To support listeners to understand speakers with a foreign accent, Bradlow and Bent's experiments (2008) revealed that native English speakers had high perceptual flexibility in adapting to foreign English accent, and after exposure to multiple talkers of Chinese-accented English during the training, native English listeners could achieve talker-independent adaption to Chinese-accented English successfully. Experiments based on the perceptual learning paradigm were conducted by Sjerps and McQueen (2010). They conducted an experiment where one group of Dutch listeners was exposed to the English /θ/ sound as a replacement for /f/ in /f/-final Dutch words, while another group heard the /θ/ sound as a replacement for /s/ in /s/-final Dutch words. As noted by Sjerps and McQueen (2010), the phonemes /s/ and /f/ are individual instances of pronunciation variation for /θ/ that can cause perceptual confusion for listeners. The result showed that the two groups of Dutch listeners could successfully interpret the /θ/ sound as /f/ or /s/. This experiment indicated that the perceptual system can adjust to variations in speech sounds and phonetic details. Connections are rapidly made between new sounds and phoneme repertoire with the aids of the perceptual flexibility. This phenomenon, known as *lexically guided retuning* (Mirman et al., 2006; Norris et al., 2003; Sjerps & McQueen, 2010), enables listeners to adapt rapidly and thoroughly to sound variability. In other words, perceptual flexibility in the perception of the phoneme representations aids the listener to adapt rapidly to accommodate idiosyncratic articulation in the speech of a particular talker.

Moreover, previous studies also applied the concept of perceptual flexibility in bilingualism and non-native language acquisition, where it plays an important role in fast learning a new language. For example, in Singh et al.'s study (2017), English was used as the native language and Hindi as the non-native language. The perceptual systems of Chinese-English bilingual infants (10-11.5 months) were found to be unconstrained by their native language's phonological repertoire. These infants did not exclusively develop sensitivity to native speech and talkers (Singh et al., 2017), demonstrating perceptual flexibility that enabled them to discriminate phonemes in both native (English) and non-native (Hindi) speech. Werker and Tees (1985) found that the early linguistic experience, that is, learning L2 at a very young age, in L2 acquisition can benefit the perceptual flexibility. English native students with early experience in learning Hindi as L2, even without further exposure to Hindi later, had no difficulty in discriminating contrastive phonemes in Hindi but which were not contrastive in English.

Therefore, the concept of perceptual flexibility indicates that the speech perception system is highly adaptable, dynamic, and actively responsive to current speech input. Recall that previous research described how perceptual flexibility facilitates sound or accent discrimination and L2 acquisition. Similar to the previous research, the *speech perceptual flexibility* applied in this thesis acknowledges that the speech perceptual system is highly flexible. However, differing from previous research, the *speech perceptual flexibility* applied in this thesis focuses on human listeners' ability to exercise flexible attentional control and make adaptable use of a range of speech units and language cues. It is a cognitive mechanism that describes listeners' cognitive ability to dynamically adjust to variations in acoustic signals and language cues during speech unit perception and sentence processing. It provides an alternative, internal, process-oriented explanation for potential differences in listeners' behavioral patterns in native and non-native sentence processing.

Specifically, this thesis explores the fascinating phenomenon of perceptual flexibility observed in the cognitive process of speech perception and sentence processing. Listeners exhibit flexible control over their attention and perception of speech perceptual elements, including segmental, suprasegmental, and sentence-level cues. The *speech perceptual*

flexibility is characterized in this thesis as a fundamental listener's strategy for handling variability in the speech signal can be operated at different level of speech perception and processing.

In the perception of acoustic features, flexibility is exemplified by *duplex perception* observed in prior studies. This phenomenon reveals listeners' ability to dynamically adjust perceptual strategies, grouping signals according to their language's phonetic inventory, even when presented separately (Ciocca & Bregman, 1989; Davis & Johnsrude, 2007; Laufer & Pratt, 2003). When the syllables *pay* and *lay* are played to each ear, listeners were able to report the fused word *play* (Cutting, 1975). Liberman and his colleagues (Mann & Liberman, 1983; Whalen & Liberman, 1987) removed the F3 from the synthetic syllables /da/ and /ga/ to make these two syllables ambiguous. They then presented this isolated formant (F3) to one ear and the rest of the syllable (the "base") to the other ear. When these two parts are presented simultaneously, although listeners report hearing the isolated F3 as a non-speech segment, their auditory system fuses the isolated formant and the "base" together. The result showed that listeners typically perceive a complete syllable, either /da/ or /ga/, depending on the F3 transition. In other words, despite the F3 and the "base" being delivered to different ears, listeners integrated them into a single, unified speech sound. This duplex perception of the fused words and ambiguous syllables underscores the idea of perceptual flexibility, demonstrating that listeners can flexibly control, combine, and organize sensory input to match their internal cognitive representations.

Moving to the segmental level, perceptual flexibility extends to phenomena like the *masking effect*, wherein listeners demonstrate flexible control over their perception of the sizes of fundamental units. In speech perception, the concept of *masking effect* highlights that listeners are quicker to perceive larger units such as syllables compared to smaller units like phonemes (Goldinger & Azuma, 2003; Norris & Cutler, 1988; Savin & Bever, 1970). However, in native speech perception, when explicitly instructed to swiftly perceive phonemes and prompted to consider phonemes as more crucial than syllables, listeners demonstrate the ability to perceive phonemes faster than syllables. In other words, listeners can consider either phoneme or syllable as the fundamental speech perceptual

unit when attention and perception are specifically directed. Listeners possess the cognitive ability to control their perception and attention, shifting across units of different sizes (Goldinger & Azuma, 2003).

At the suprasegmental and sentence levels, perceptual flexibility continues to manifest through the utilization and integration of cues. Native and non-native listeners show a tendency to prioritize certain cues over others. They also display flexibility in allocating attention weights to and integrating specific cues. For instance, non-native listeners may prioritize the utilization and integration of contextual and semantic cues while possibly disregarding prosodic and syntactic information (Akker & Cutler, 2003; Clahsen & Felser, 2006b; Ganga et al., 2024; Juffs & Harrington, 1995; Vanlancker–Sidtis, 2003; Ying, 1996).

Therefore, the *perceptual flexibility* applied in this thesis encompasses the perception, utilization and integration of the segmental, suprasegmental, and sentence-level cues. This speech perceptual flexibility operated at two levels: speech unit perception level (perception) and sentence processing level (comprehension).

Speech perception can be challenging in various contexts. Native listeners demonstrate remarkable adaptability in their ability to quickly and effortlessly compensate for large acoustic changes and signal degradation (Cutler, 2012; Pisoni, 2018), unlike non-native listeners who encounter numerous difficulties during real-time speech perception and comprehension (Bohn, 2018; Hopp, 2015; Patterson et al., 2017; Segalowitz, 2003). Therefore, L2 listeners often encounter difficulties, potentially stemming from their inability to fully leverage perceptual flexibility to dynamically react to incoming signals.

To address the central question in speech perception and comprehension–"what makes non-native speech and language difficult to perceive and process?" (Bohn, 2018)–I examine whether the difficulties that listeners encounter in non-native speech perception are due, in part or wholly, to an inability to fully utilize their perceptual flexibility to dynamically respond to incoming signals. Specifically, this thesis investigates the speech perceptual flexibility exhibited through the different strategies and abilities demonstrated

by both native and non-native listeners in low-level speech unit perception and high-level sentence processing.

Thus, my focus lies in exploring perceptual flexibility under the context of native and non-native speech perception and sentence processing. Generally, perceptual flexibility operates at two levels: (1) I conducted phoneme and syllable monitoring experiments to investigate listeners' attention shifts between the phoneme-size and syllable-size units when dealing with speech unit perception; (2) I conducted eye-tracking experiments to examine listeners' allocation of attention to multiple cues during sentence processing, including pitch accent as a suprasegmental cue in relation with information structure and verb semantics as sentence-level cues.

Specifically, the phoneme and syllable monitoring experiments are designed to address the following research questions: Do changes in either bottom-up or top-down information, or simultaneous changes in both, affect the masking effect for native and non-native listeners? I predict that non-native listeners lack the same degree of perceptual flexibility as native listeners at the level of low-level speech unit perception. If this is correct, the monitoring experiments will show more significant changes in the magnitude of the masking effect for native listeners compared to non-native listeners.

Further, the eye-tracking experiments aim to answer the research questions: do non-native listeners rely more on verb semantic information than prosodic pitch accent information? Are non-native listeners less able to combine semantic and prosodic pitch accent information interactively compared to native listeners? I predict that non-native listeners have a reduced ability to utilize and integrate multiple cues in high-level sentence processing compared to native listeners. If this is correct, the eye-tracking experiments will show that non-native listeners are less attentive to pitch accent than to verb semantic information and are also less able to combine the three types of cues interactively.

1.2 Outline of chapters

This thesis comprises monitoring and eye-tracking experiments, each examining the *perceptual flexibility* of listeners in both native and non-native speech perception. These experiments address two aspects: listeners' ability to shift attention between fundamental speech perceptual units at low-level speech perception and their allocation of attention weight to multiple cues for high-level sentence processing.

Chapter 2 This chapter thoroughly introduces the research background for the monitoring experiments. This chapter introduces essential concepts related to the mechanisms that reveal the perceptual flexibility in speech perception. Phoneme and syllable are the main candidates for fundamental perceptual units. It also illustrates the phenomenon of masking effect in perceiving these fundamental speech units. Fluctuation in the size of the masking effect serves as an indicator of listeners' capacity to shift attention among these units. This chapter offers a detailed exploration of speech perceptual models, with a specific emphasis on the mechanisms of perceptual flexibility governing the representation and retrieval of speech units. Particularly, it illustrates how memory episodes in exemplar models are activated based on stored auditory properties and their corresponding attention weights. Further, the chapter discusses the formulas in MINERVA 2, which could depict the cognitive processing of both native and non-native listeners during real-time speech perception.

Chapter 3 This chapter introduces the experimental methods for the monitoring experiments (Experiments 1 and 2). To investigate the perceptual flexibility operating at the level of speech perception, phoneme and syllable monitoring experiments were conducted in two parts with native and L2-speaking English and Mandarin listeners listening to English pseudo-word stimuli (Experiment 1) and Mandarin pseudo-word stimuli (Experiment 2). Listeners' attention was selectively directed towards phoneme-level information through one of two manipulations. First, an artificial accent, in which the pronunciation of specific phonemes was altered, served as bottom-up information directing attention to the phoneme level. Second, explicit instructions as top-down information about the nature of the artificial accent were provided before the experiment began. In this way, both bottom-

up information (accent) and top-down information (explicit instruction as prior knowledge on the artificial accent) were made available to listeners. These manipulations therefore allowed me to examine listeners' ability to use bottom-up, top-down, or both types of information to shift their attention across perceptual units of different sizes. Listeners listened to pseudo-words that started with the target phonemes or syllables displayed on the screen. Listeners' attention shifting across speech units is demonstrated through the variations in the magnitude of the masking effect, which reveals perceptual flexibility and sheds light on the ways in which perceptual units are stored and retrieved during real-time native and non-native speech perception.

Chapter 4 This chapter presents and discusses the results of Experiments 1 and 2. Both native and non-native listeners exhibited delayed reactions to phoneme and syllable targets when the targets were presented with an artificial accent, compared to when the targets were presented without an artificial accent. Accent as bottom-up information plays a more significant role than prior knowledge as top-down information in directing listeners' attention across speech units. Further, the phoneme and syllable monitoring experiments unveiled potential language-specific differences in perception, representation and retrieval of perceptual units. Under the Mandarin context, Mandarin native listeners showed no significant difference observed in the susceptibility of phoneme and syllable units to bottom-up, top-down, or combined information influences. But under the English context, both Mandarin and English listeners showed more pronounced changes in response to syllable targets compared to phoneme targets. This indicates that syllable processing could be more susceptible to disruption than phoneme processing, further suggesting that, especially in English, phonemes could be a more robust speech unit than syllables during speech perception.

Chapter 5 This chapter shifts focus from perceptual flexibility at the segmental level to perceptual flexibility at the sentence level. It presents an overview of the utilization and integration of cues, with a particular focus on the abilities of native and non-native listeners to use and integrate multiple cues during real-time sentence processing. The chapter begins by examining various theories and hypotheses that explain the differences in per-

formance between native and non-native listeners in processing sentences. Following this theoretical foundation, the chapter proceeds to review previous research investigating the disparities in how native and non-native listeners utilize and integrate semantic and prosodic information during real-time sentence processing.

Chapter 6 This chapter introduces the experimental methods for the eye-tracking experiments (Experiments 3 and 4). To investigate perceptual flexibility in sentence processing, Experiments 3 and 4 were again conducted in two parts with native and L2-speaking English and Mandarin listeners. In Experiment 3, participants listened to English sentences, while in Experiment 4, they listened to Mandarin sentences. This study aimed to determine the cues that listeners rely on most and to explore their perceptual strategies during real-time sentence processing. In the experiments, participants listened to sentences presented within a brief discourse context. The sentences varied according to three binary factors: prosodic pitch accent (neutral or contrastive pitch accent), target object information structure (new or repeated information), and semantic match between the verb and the target object (appropriate or inappropriate verb semantics). Participants were instructed to click on the object named in the second sentence that was presented on the screen, while their eye movements were recorded.

Chapter 7 This chapter presents and discusses the results of Experiments 3 and 4. Results showed a complex interaction between L1, prosody, and verb semantics. When processing sentences with repeated information, both native and non-native listeners in English and Mandarin contexts demonstrated the ability to utilize prosodic and semantic information, along with exhibiting an interaction between the prosodic and semantic information. However, when processing sentences with new information, non-native listeners exhibited a reduced ability to utilize and integrate different cues, suggesting that they can only adopt native-like strategies in integrating prosodic and semantic cues when the information structure of target words is familiar, thus making it less taxing to process. Further, the utilization and integration of multiple cues for sentences containing new information differed between English and Mandarin contexts.

Chapter 8 This chapter summarizes the results of both monitoring and eye-tracking experiments in relation to the objectives of this thesis, illustrating how perceptual flexibility is reflected at both low-level speech unit perception and high-level sentence processing. The discussion focuses on how this flexibility of listeners manifests through the representation and retrieval of perceptual units, as well as the utilization and integration of cues, shedding light on the effortless nature of native speech perception and the challenges associated with non-native speech perception. Further, the chapter discusses the limitations of the study and points towards future research avenues highlighted by this thesis.

Chapter 2

Perceptual flexibility of unit perception: attention shifting between phoneme and syllable

2.1 Speech perceptual units

Speech perceptual units are a widely held theoretical concept defining the minimal sound patterns assumed to exist in continuous speech, thus constituting the speech signal as a sequence of these units. Each perceptual unit contains information consisting of a set of acoustic features stored in long-term memory (Massaro, 1974). The concept of speech perceptual units serves theoretical purposes from two facets.

Firstly, speech perceptual units theoretically provide an explanation of how listeners and speakers translate continuous physical signals into abstract representations. Liberman and Whalen's (2000) interpretation of speech perceptual units was symbolic, differing from the exemplar theoretic view, which posits that perceptual units are stored as memory traces, such as that of Hintzman (1984, 1986, 1988). Liberman and Whalen (2000) suggested that speech perceptual units are artifacts, akin to "the letters of the alphabet." Rather than being inherent in physical speech signals, these units, such as phonemes and syllables, embody abstract symbols essential for formalizing the structure of speech within linguistics. Human speech perception entails a sequence of "abstract, idealized, context-free symbols" (Pisoni, 2018). This sequence typically comprises phonemes and syllables arranged in a linear order, aligning with the traditional abstractionist perspective in linguistics. Hockett (1956) analogized the generation of speech involving the

co-articulation between speech perceptual units to the process of smashing a row of Easter eggs, which effectively captures the complexity and co-articulation of these units in speech production. Building on Hockett's ideas, Liberman and his colleagues proposed the concept of "parity" (A. Liberman & Whalen, 2000; Mattingly, 1990) which serves as a crucial agreement between speakers and listeners on what sounds constitute meaningful speech elements. The perceptual units are the Easter eggs arranged in a row along a moving conveyor belt. Then this row of eggs is smashed by a wringer, resulting in a mess containing the materials of the smashed eggs, which represents a mixture of output of spoken language. This metaphor effectively conveys the intricacy of how these units blend together in speech. In the context of the Easter egg analogy, the smashing process allows distinct speech units (the eggs) to be transmitted from speaker to listener, but not in an easily recognizable form. This process fulfills the requirement of "parity," which underscores, despite the chaotic process of co-articulation, there is an underlying agreement on the meaningful arrangement of speech perceptual units, allowing for effective communication. Speakers and listeners have this deliberate agreement and common recognition regarding what constitutes meaningful speech perceptual units. Similarly, Liberman and Whalen (2000) described this common recognition, stating that people have to know that the sound, such as /ba/, counts as a phonetically meaningful speech perceptual unit, while a sniff or other meaningless noise segments do not. Therefore, speech perceptual units serve as foundational constructs that enable the interpretation of acoustic signals through a shared understanding among speakers and listeners. This consensus on meaningful units underscores the collective recognition of specific sounds as significant segments while excluding irrelevant noises, thus facilitating effective communication.

Secondly, due to the characteristics of speech perception, for the sake of economy and convenience, a certain length of segments in acoustic information should be stored before coding commences. Miller (1962) referred to these segments as "decision units" in speech perception. Psychologists proposed a "slow and single-channel" decision-making mechanism in human cognition, implying that input information should be stored and processed as a unit, leading to decisions being made at specific times to determine unit boundaries. An early study by Shannon (1948) noted that sequences in a language do

not appear randomly; for instance, in English, the letter E has a higher frequency of occurrence than the letter Q, and the sequence TH occurs more frequently than XP. Thus, it is reasonable to predict that certain length sequences are already stored in cognition to save time in coding, and these stored segments allow people to encode message sequences into signal sequences (Shannon, 1948). For theoretical purposes, Miller (1962) suggested that it would be convenient for listeners if the length of stored segments were infinite. However, in practice, only a finite length of coding unit, such as a word, is allowed to be stored. Consequently, practical limitations in cognitive processing and representation necessitate the representation of acoustic information in distinct units to optimize speech perception and coding. Building upon this foundation, endless messages are constructed from small sets of units. Goldinger and Azuma (2003) characterised this phenomenon as "unitization," which is logically required in speech processing due to the hierarchical structure of language. Unitization has been also introduced in earlier research (Goldstone, 1998; Samuel & Kraljic, 2009); as one of the mechanisms in perceptual learning, referring to the ability to distinguish critical units from the whole. Thus, given the unique characteristics of speech processing, listeners must retrieve units from continuous speech signals to understand speech.

Determining the fundamental unit of perceptual analysis has been a longstanding research question. Previous studies exploring how listeners perceive these fundamental units have revealed their ability to discern speech perceptual units of different sizes. Although the results have been fruitful, they have also presented inconsistencies. While phonemes and syllables are conventionally considered two potential candidates for perceptual units in speech, some research suggests that these fundamental units could be larger entities such as words (Klatt, 1979; Miller, 1962) and phrases (Miller, 1962). For example, early work by Miller (1962), suggests that phrases, consisting of two or three words, may serve as potential natural decision units in speech perception. Moreover, Liberman et al. (1967) suggested that CV syllable was perceived as an entire unit since the stop consonant /d/ cannot be perceived independently of perceiving a CV syllable. Savin and Bever (1970) argued that phonemes, rather than being perceptual, are abstract and are identified after syllables. Further, Foss (1971) conducted experiments where listeners were tasked with

identifying either syllable targets or word targets from lists of words. The results indicated that responses to words were quicker than responses to syllables, suggesting that words could also function as perceptual units in speech.

Nevertheless, the debate regarding the fundamental unit predominantly centers on phonemes and syllables as the primary candidates for speech perceptual units (Decoene, 1993; Goldinger & Azuma, 2003; Massaro, 1975; Norris & Cutler, 1988). This thesis aligns with the perspective that phoneme and syllable are viewed as candidates of the fundamental units in speech perception. The prioritization of either phonemes or syllables for listeners during speech perception is a stance that is further investigated through experimental studies in this thesis.

2.2 Masking effect as a phenomenon

Linked to the fundamental speech perceptual units outlined in the previous section, in the strictly hierarchical models of speech perception, it is anticipated that smaller units such as phonemes would be identified more rapidly than larger units like syllables, words, and phrases (Norris & Cutler, 1988; D. A. Swinney & Prather, 1980). However, as described in Chapter 1, larger units can be identified more rapidly than smaller units during speech perception, which has been extensively documented (Goldinger & Azuma, 2003; Norris & Cutler, 1988; Savin & Bever, 1970). This robust tendency has been termed the *masking effect*. During regular language usage, the focus of listeners predominantly lies on larger units like words, optimizing the transmission of information. Thus, larger units typically "win" in listeners' identification during speech perception, creating a masking effect.

Foss and Swinney (1973) explained why phoneme as smaller units would gain slower reaction by noting that experiments on phonemes and syllables reveal more about the order of identification than the order of perception. Consequently, it is plausible that syllables are identified more quickly than phonemes. This does not negate the possibility that phonemes may be the primary or initial units of perception; however, they might not be identified as readily as larger units.

Among the early studies demonstrating that listeners respond more quickly to larger units than to smaller ones, for example, in Savin and Bever's study (1970), they asked participants to responded as soon as they heard a preset target in a sequence of nonsense syllable. The target was a complete syllable (e.g. /bæb/), the syllable-initial consonant phoneme from some objects (e.g. /b-/, /s-/), or the medial vowel phoneme(e.g. /-æ-/). The results showed that responses to initial phonemes were slower compared to responses to entire syllables. Similarly, in study of Segui et al. (1981), participants were asked to identify either the initial stop consonants (e.g. /b/, /d/, /p/) or the initial CV syllables of bi-syllabic words and non-words (e.g. /ba/, /de/, /pi/) presented in mixed lists. The results also showed that syllable detection was consistently faster than phoneme detection, suggesting that syllable identification precedes phoneme recognition. These studies served to illustrate how larger linguistic units mask smaller ones, confirming the existence of the masking effect as a phenomenon.

This thesis defines the masking effect as an indicator of perceptual flexibility, closely associated with the *adaptive resonance* proposed in Adaptive Resonance Theory (ART) (Carpenter & Grossberg, 1987; Goldinger & Azuma, 2003; Grossberg, 1999; Grossberg, 1976a, 1976b, 1980, 2003, 2013). The adaptive resonance proposed in ART is a dynamic brain state achieved through the interaction between bottom-up (input signals in working memory) and top-down information (prior knowledge stored in long-term memory) during speech perception (see detailed illustration of adaptive resonance in Section 2.3.2.2 *Models with Adaptive Resonance*). When adaptive resonance is achieved, listeners' attention is drawn to a specific unit, which could be either a phoneme or a syllable. As mentioned previously, typically the larger units such as syllables "win" listeners' attention and are identified faster than phonemes. However, by manipulating the bottom-up and top-down information to achieve listeners' adaptive resonance and direct their attention to phonemes, the phenomenon of *reversed masking effect* can be achieved.

Goldinger and Azuma (2003) demonstrated this *reversed masking effect* in an experiment where, to manipulate bottom-up information (physical acoustic-phonetic cues), the speakers of the stimuli were informed by stories that either phonemes or syllables are

fundamental units. These stories led phoneme-biased speakers lengthened fricatives and accentuated stop releases to create phoneme-biased stimuli, while syllable-biased speakers amplified prosodic cues and omitted initial consonants to produce syllable-focused stimuli. Correspondingly, the top-down information (listener's expectations) was also manipulated by informing listeners of the primacy of phonemes or syllables in speech perception, thus biasing their perceptual expectations towards either phonemes or syllables. As a result, they interpreted these findings as showing that when the listeners were informed with phoneme-biased top-down information and given with the phoneme-biased stimuli, listeners demonstrated faster responses to phonemes over syllables in their native speech perception (Goldinger & Azuma, 2003). The presence of strong phonetic cues to smaller units, combined with prior knowledge or expectation about the primacy of these smaller units, can overcome the usual pattern where larger units mask smaller ones. This allows these smaller units to be perceived faster than their carrier syllables or words. As human listeners naturally tend to react faster to larger units like syllables than smaller units like phonemes, deliberately directing attention to phonemes should enable the focus on units not typically attended to, showcasing perceptual flexibility in flexibly and dynamically directing attention to an unconventional level of units.

Therefore, the fluctuation in the size of the masking effect is an indicator of perceptual flexibility, as the phenomena of masking effect and reversed masking effect highlight the flexibility of speech perception and the complex interplay between cognitive processes and linguistic units. If the masking effect fluctuates, it means that listeners' attention across phoneme and syllable varies, with either the phoneme or syllable "winning" more attention, resulting in changes in the magnitude of the masking effect. If the phoneme "wins" more attention, in other words, the adaptive resonance is achieved on the phoneme, a reversed masking effect is achieved, as demonstrated in Goldinger and Azuma's study (2003).

In this thesis, I conducted monitoring experiments (Experiments 1 and 2) revolving around the masking effect as a central concept. The masking effect serves as an indicator of the dominant perceptual unit, shedding light on the adaptable nature of human speech per-

ception. This adaptability is central to understanding how larger linguistic units, like syllables, can often overshadow smaller ones, such as phonemes, in the listener's perceptual field. However, this overshadowing is not absolute. This flexibility, where the balance between the recognition of phonemes and syllables can shift, highlights the adaptable and malleable nature of human cognitive processes during speech perception.

The masking effect can be quantified by measuring the discrepancy in RTs between phoneme and syllable recognition. The methodology employed to calculate and quantify the masking effect for the monitoring experiment is detailed in Chapter 4 (see Section 4.1.2 *Calculation of masking effect*).

2.3 Models for speech perceptual units processing

2.3.1 Two key issues: activation and representation

Speech perception traditionally involves correlating speech signals with mental representations, yet historical theories have suggested that understanding speech involves decoding the actions of the vocal tract, which conveys acoustic information and constitutes the process of speech signals. Consequently, theories on phonetic perception diverge into two main theoretical categories. The first posits that the actions of the vocal tract correspond to phonological categories (Goldstein & Fowler, 2003; A. M. Liberman & Mattingly, 1985), represented by models such as Analysis-by-Synthesis (AxS) model (Bever & Poeppel, 2010) and the Direct Realist model (Fowler, 1986), advocating for a direct link between auditory input and gestural information in speech perception. The second theoretical category contends that listeners identify cues in the acoustic speech signal and then align these cues with their mental phonological categories (Diehl & Kluender, 1989; Sawusch & Gagnon, 1995).

In this thesis, I examine models categorized under the second theoretical approach, which establishes a linkage between acoustic-phonetic signals and mental representations. This perspective fundamentally posits that listeners detect cues within the acoustic speech signal and subsequently match these cues with their internal mental phonological categories. This viewpoint is exemplified by by Diehl and Kluender (1989) as well as Sawusch and

Gagnon (1995): the long-standing fundamental presumption of speech perception is that the speech perception depends on the recognition of abstract, idealized, and context-free symbolic representations. Goldinger (2007) has also observed that within cognitive science, two core concepts guiding theory and research are, firstly, that perception and categorization involve stable mental representations, spanning from individual memories to universally shared knowledge, and secondly, these representations are activated by transient operations. Consequently, mental representations are a crucial aspect. The current models discussed in this section focus on how speech signals are mapped onto the mental representations in the process of speech perception.

In the context of speech perception as a mental mapping process, Dahan and Magnuson (2006) addressed two critical issues in speech perception theories, highlighting differences in how current models approach these two important aspects for the speech perceptual unit: activation and representation. By addressing both speech perceptual processing and lexical access, these models are classified into different groups based on their approaches to these two critical aspects.

The first critical issue is unit activation, which involves two key interacting elements: the activation of fundamental units and the flow of information. Traditional models, such as TRACE (McClelland & Elman, 1986) and the Cohort model (Marslen-Wilson, 1987) (see also Section 2.3.2.1, *Models with Fixed Hierarchical Activation*), usually suggest that the units activated during speech perception are of fixed sizes, such as acoustic features, phonemes, syllables, and words, etc., whereas more dynamic models propose the concept of *adaptive resonance*.

The adaptive resonance (or cognitive resonance), emerges when bottom-up information and top-down expectancy coalesce. In other words, adaptive resonance is the resonant brain state. Bottom-up information and top-down expectancy are the two factors that produce the resonant brain state. This state can only be achieved through the interaction between bottom-up input and top-down prior expectancy (Goldinger & Azuma, 2003, 2004), which then draws listeners' attention and creates a conscious experience. Therefore, the

model of speech perception, such as Adaptive Resonance Theory (ART) (Goldinger & Azuma, 2003; Grossberg, 1976a, 1976b), incorporates the concept of perceptual units, such as phonemes and syllables, but argues that the perception of unit sizes is flexible and determined by listeners' attention during the interaction between bottom-up and top-down information. The entity (e.g. phoneme, syllable) that gains listeners' attention is considered the perceptual unit. This is similar to the statement by McNeill and Lindig (1973), who noted that *perceptual reality* is what one pays attention to. Extensive research also indicates that native listeners are adept at identifying perceptual units of varying sizes as the primary units of speech perception (Goldinger & Azuma, 2003; Healy & Cutting, 1976; D. A. Swinney & Prather, 1980), which could further support that the perceptual reality is what one pays attention to.

Now consider the information flow. Speech perception models allow information to flow in one of two ways: either autonomously, or interactively. From the autonomous view, the information flow allows only for feedforward processing. This autonomous perspective contends that optimizing processing involves permitting only bottom-up information, disallowing unnecessary interaction with top-down information, which doesn't influence bottom-up information flow (Norris et al., 2000, 2003). Dahan and Magnuson (2006) explained the autonomous processing and introduced the concept of "veridical perception," suggesting that the integration of top-down information with sensory input poses a challenge in distinguishing between environmental and organismic information, therefore the suggested that the top-down information integration is unnecessary. Norris et al. (2000) argued that top-down information is never necessary during speech perception. They provided several reasons for this claim, including the argument that feedback from words in long-term memory to facilitate the recognition of sound inputs violates the principle of simplicity (Occam's razor) because such feedback introduces additional, unnecessary load and complexity in speech perception. They also pointed out that top-down feedback does not improve word recognition. Listeners were able to detect phonemes, restore phonemes, and assign spoken input to phoneme categories without relying on top-down lexical and contextual feedback.

By contrast, other models argue that speech perception can involve interaction between representational levels, that is, speech perception can also be interactive. The integration of bottom-up and top-down information during speech perception is plausible. Resolving the veridical perception issue involves assigning proper weights to both bottom-up and top-down information, enabling a bidirectional information flow. In speech perception, top-down information plays an important role. Top-down information can feed back to contribute to the predictive process (Corps & Rabagliati, 2020; Sohoglu et al., 2012). Additionally, more dynamic models and theories such as MINERVA (Hintzman, 1984, 1986, 1988) and ART (Carpenter & Grossberg, 1987; Grossberg, 1999; Grossberg, 1980, 2003, 2013), also emphasize the interplay within information flow and highlight the potential for an interactive process that integrates both bottom-up and top-down information.

Together, in the context of unit activation, both the activation of fundamental units and the flow of information stand as crucial systems and mechanisms. These mechanisms are therefore inherent components of speech models during the unit activation process. Generally, models that posit fixed units for activation can incorporate a flow of information system that is either autonomous or interactive. The more dynamic models, which propose adaptive resonance to capture listeners' attention for achieving the same function and effect as the "unit" in speech perception, typically adopt an assumption of interactive information flow.

The second critical aspect under discussion is the manner in which perceptual units are represented. Two categories of models can be distinguished: the prototype models, which argue that units are stored as idealized prototypes, or abstract forms; and the exemplar models, which argue that units are stored as much more concrete representations (Weber & Scharenborg, 2012). Frixione and Lieto (2012) gave the example that the concept of animal CAT should coincide with a representation of a prototypical cat, while in the exemplar view, the concept of animal CAT is a set of representations of the cats we encountered during our lifetime.

In speech, prototype models advocate for the activation of units based on abstracted rep-

resentations, known as "phonetic prototype" (Oden & Massaro, 1978; Samuel, 1982). These models emphasize the acoustic input's journey through various modifications, re-encodings, and a process of normalization. Normalization enables listeners to identify the common phonetic elements across different realizations of phonemes, facilitating the quick understanding of speech from any speaker, even if they have not been encountered before (McQueen & Cutler, 2010). Thus, phonetic processing is heavily reliant on a prototype-based system, which forms the foundation of what is known as the "conventional symbol-processing approach" to speech perception. This conventional symbol-processing approach treats the variability as a source of noise in speech signals that should be eliminated to uncover the hidden abstract underlying symbolic linguistic messages (Elman & McClelland, 1986).

In contrast, exemplar models, such as Hintzman's MINERVA model (Hintzman, 1984, 1986, 1988), propose that psychological lexical representations are detailed episodic memories that include fine acoustic details of speech, stored within long-term memory. These models highlight the significance of individual speech events and the rich acoustic details they contain, suggesting that our understanding and recognition of speech are grounded in these detailed episodic memories rather than abstracted prototypes (Goldinger, 1998; Hintzman, 1984). Therefore, the exemplar theory distinguishes itself by adopting a "non-analytic approach" to speech perception, emphasizing the encoding of specific instances of perceptual events. This focus on detailed episodic memories showcases the exemplar theory's unique perspective (Pisoni, 2018).

Taken together, reviews of speech perception models in the following sections will focus on two key issues: activation and representation. Models with a fixed hierarchical mechanism also debate the information flow, specifically the timing of interaction. In contrast, by acknowledging the initial integration of top-down information, models without a fixed hierarchical mechanism typically propose an interactive view of information flow. The monitoring experiment in this thesis was conducted to reveal the flexibility in perceiving units and to determine the extent to which bottom-up and top-down information influence listeners' perception. This experiment aims to provide new experimental evidence for

models of speech perception.

2.3.2 Information flow and activation

2.3.2.1 Models with fixed hierarchical activation

As outlined in the preceding section, there are two categories of speech perceptual models regarding the unit activation: those that posit the activation of fixed units in a hierarchical order and those that focus on adaptive resonance. As illustrated in the preceding section, adaptive resonance, which also refers to resonant brain states, is the product of the interaction between bottom-up input and the top-down expectancy of listeners. Classical models with multilevel structures for unit activation propose the activation of fixed units from smaller to larger scales. Upon hearing a word, the process activates multiple lexical representations, determined by their frequency and how closely they align with the incoming speech input. The recognition phase is then shaped by a competitive dynamic among these representations. The classic models are exemplified by TRACE (McClelland & Elman, 1986), the Cohort model (Marslen-Wilson, 1987), and the Fuzzy Logical Model of Perception (FLMP) (Massaro, 1989; Massaro & Chen, 2008).

These models conceptualize speech perceptual units in a concrete, fixed, and hierarchical fashion, depicting them as nodes spanning from phonetic feature layers to word layers. Activation typically proceeds from lower to higher layers, illustrating a structured information flow that underscores the presumed fixed hierarchy of speech perceptual units. Within the range of models that propose a fixed hierarchical activation of units, these models differ in exemplifying the mechanism of information flow. For example, a typical hierarchical order in TRACE model is illustrated in Figure 2.1.

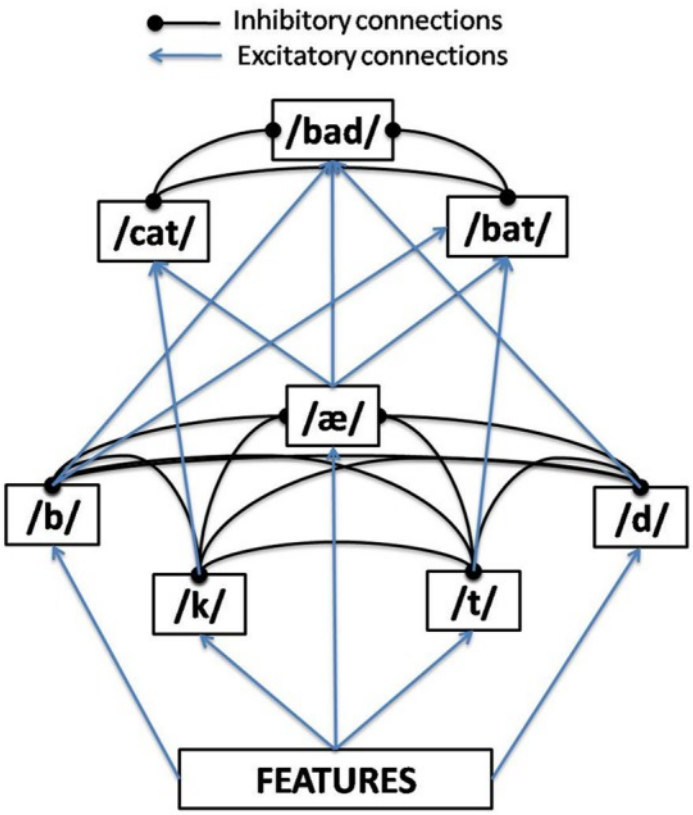


Figure 2.1: Typical hierarchical structure in TRACE model. Reprinted from Hannagan et al (2013)

As for the issue of information flow, as stated in the preceding section, the influence of bottom-up information, such as the quality of suprasegmental cues, the ambiguity of different phonemes, and variations in testing methods, is undeniably crucial in determining the detection time of different units (Goldinger & Azuma, 2003). Similarly, top-down information also plays a significant role in real-time speech perception. Twilley and Dixon (2000) identified pre-existing knowledge, such as expectations, contextual information, surrounding words, and prior knowledge, as top-down feedback for interpreting new sensory information. This contrasts with new sensory and perceptual information, considered bottom-up input to be mapped onto lexical representations.

Generally, all models of speech perception necessitate some form of feedback to integrate contextual and perceptual information, thereby achieving final meaning resolution (Magnuson et al., 2018; Twilley & Dixon, 2000). These models, which predict a fixed hierarchical representation unit, recognize the crucial interaction between bottom-up input and top-down confirmation required by human perceptual system. A fundamental theoretical perspective in cognitive science is the *interactive activation hypothesis*, proposing that forward and backward flows in bidirectionally connected neural networks enable the cognitive system to integrate bottom-up and top-down information effectively (Magnuson et al., 2018).

Therefore, among these speech perceptual models positing the fixed hierarchical units, there exists a feedback interaction between higher and lower levels of representation, with bidirectional information flow providing a viable solution. This interactive systems activation can tune the perceptual systems to prior knowledge based on the experience. Further, feedback from this prior knowledge offers an efficient means to resolve ambiguous and obscured bottom-up information (McClelland & Rumelhart, 1981; McClelland et al., 2014). Consequently, speech perception is conceptualized as a cognitive process that involves the dynamic interplay between bottom-up and top-down information flows. The influence of top-down information, encompassing prior knowledge and context effects, is well-documented across numerous studies. For instance, phenomena such as phoneme restoration illustrate that listeners can perceive a segment as present even when

it has been entirely replaced by noise, demonstrating that higher-level lexical knowledge can provide feedback to phonemes at a lower level, resolving vagueness or ambiguity (R. M. Warren, 1970). Additionally, studies on lexical semantic priming and target detection have showed that contextual information, serving as top-down feedback, can either facilitate or inhibit lexical processing and target detection within sentences. For example, individuals respond faster and more accurately to a target word when it is preceded by a related word, compared to an unrelated one (Balota & Paul, 1996; Hodgson, 1991; Neely, 1990). Moreover, a congruent context or related words can decrease the time required to recognize a target word, whereas an incongruent context can extend it (Stanovich & West, 1983). Therefore, the impact of top-down feedback, including prior lexical and contextual knowledge, is pervasive across all models of language processing, which incorporates some form of feedback mechanism to resolve the ambiguity occurring at the levels of processing individual letters, words, and discourse.

Thus, the debate concerning information flow in models that predict hierarchical fixed units is not about whether top-down information is utilized during real-time speech perception, but rather about the timing of top-down influences. Essentially, contemporary speech models recognize the integration of top-down information and its contextual influence on the outcome of linguistic processing; however, there is a divergence of views regarding when this integration takes place. Consequently, the primary distinction among current models lies in their approach to information flow across the processing system's levels (Weber & Scharenborg, 2012). Specifically, the question arises: when does the top-down influence occur (Twilley & Dixon, 2000)?

Therefore, the present models can be broadly classified into two categories based on their perspective on the timing of top-down influences in information flow: those adopting an interactive approach (non-modular approach) versus those favoring an autonomous approach (modular approach). These two approaches differed in two processes: the access process and integration process. The access process refers to the act of accessing lexical representations, namely, mapping the input signal information onto these representations. The integration process involves combining prior knowledge with the lexical representa-

tions.

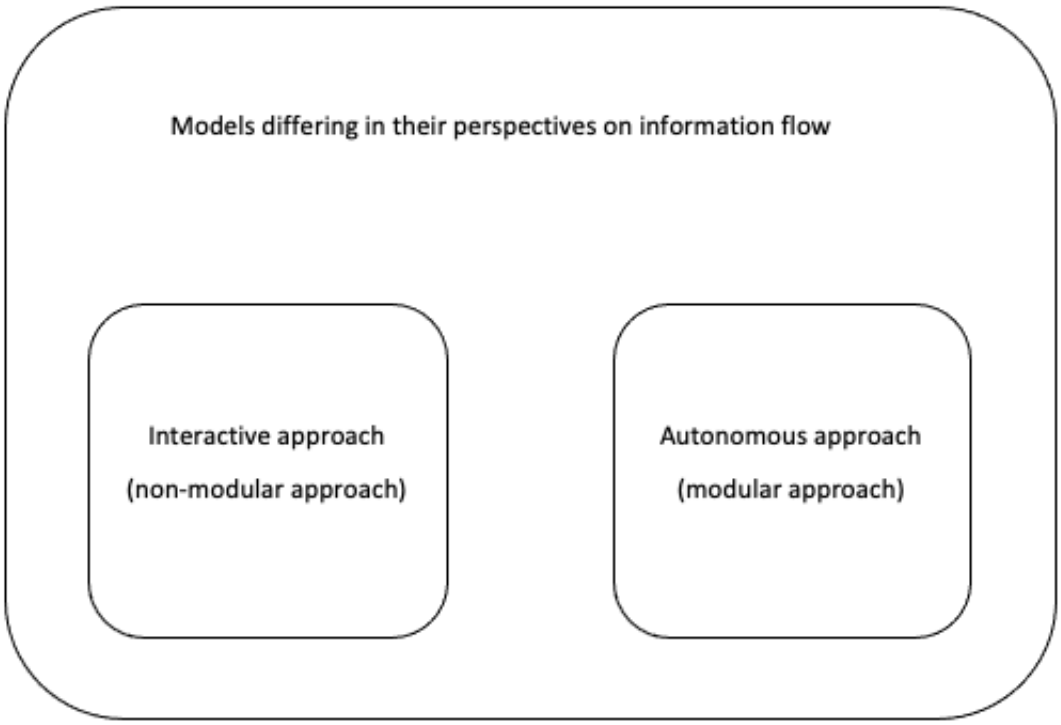


Figure 2.2: Models differing in their perspectives on information flow

The interactive view, or the non-modular view, suggests that the access and integration processes occur in parallel and interact continuously throughout the course of meaning recognition and activation. This non-modular approach posits that information flow can oscillate back and forth between two or more processes and levels. In this dynamic interaction, high-level constructs, representing prior knowledge, form direct connections with low-level segments to provide feedback, which can either facilitate or impede the activation of these lower segments. Top-down information thus flows back to the lower regions or levels, where contextual and perceptual information are synthesized to form the final decision on meaning. Models predicting the hierarchical activation about the organization of linguistic units vary in their approaches to information flow.

For example, the TRACE model (McClelland & Elman, 1986) stands out as a representative interactive, or non-modular model (Twilley & Dixon, 2000; Wayne S., 2000), which posits top-down influences from higher levels can affect activation alongside bottom-up sources, and this top-down influences occurs at the initial stages of speech and lexical

processing. TRACE is a connectionist model, where phonetic features, phonemes, and words form three interacting layers of nodes. The first layer corresponds to phonetic features, the second to phonemes, and the third to words, offering a coherent framework for integrating multiple bottom-up information sources in speech perception. Upon receiving input signals, nodes are activated in accordance to their resemblance to the input. Initially, feature nodes are activated; this activation then spreads to phoneme nodes and ultimately to word nodes, activating the three layers in a hierarchical manner (McClelland & Rumelhart, 1981; McClelland et al., 2014; Norris et al., 2000; Weber & Scharenborg, 2012).

Bottom-up processing facilitates the activation of lower-order units, such as acoustic phonemes, which in turn can initiate the activation of higher-order units. For instance, recognizing the acoustic phoneme /k/ can lead to the activation of words containing the /k/ phoneme, thereby enhancing word recognition based on phonetic inputs. Consider the word *CAT*. In a bottom-up processing scenario, hearing the initial phoneme /k/ in *CAT* provides auditory information that sequentially activates the corresponding phonemes /k/, /æ/ and /t/. This activation moves from the auditory perception of the individual sounds to the recognition of the whole word, *CAT*.

Conversely, the model also suggests that the activation of higher-order units can influence their lower-order counterparts. Specifically, the activation of a word that includes a /k/ phoneme could subsequently activate the /k/ phoneme itself. In a top-down processing scenario, context plays a critical role. For instance, upon hearing the sentence *The mouse was stopped by the sharp eyes of a hungry...*, the semantic context allows listeners to predict the word *CAT* before hearing the full word. This prediction is based on listeners' knowledge and expectations about lions and where a thorn might be. As a result, the mental representation of *CAT* is activated first, which then primes the recognition of the initial phoneme /k/ when listeners finally hear it. In this way, bottom-up processing relies on sequential auditory information to build up to the word *CAT*, while top-down processing uses contextual clues to predict and then confirm the word, leading to activation that spreads from the whole word *CAT* back to its initial sound /k/.

This bidirectional activation process underscores the dynamic interplay between higher-order and lower-order units in speech perception. The mechanisms of activation and interaction among levels are pivotal in the TRACE model, which is centered around interactive activation. It underlines that sensitivity to one source of information can be modulated by another, highlighting that the TRACE model's provision for the early integration of top-down and bottom-up information classify it as a non-modular model.

In contrast to models that adopt an interactive approach, the autonomous information flow perspective argues that online feedback is not beneficial. The autonomous view posits that the access and integration processes occur simultaneously but independently. In this perspective, the access is driven solely by perceptual input, without interference from higher-level cognitive processes such as contextual information. Instead, perceptual and contextual information are combined in a separate decision module, also known as "independent activation" (Twilley & Dixon, 2000).

According to this view, the integration process does not engage in the initial stage of lexical mapping but takes place after the access process to finalize the decision. Therefore, Twilley and Dixon (2000) commented that within the modular approach, top-down influences contribute to the "post-access processing," rather than the initial stage of the lexical perception. Within this framework, the integration of contextual and perceptual information occurs at a distinct decision selection level (Mirman, 2008). This view advocates for a representational organization that is modular and autonomous, where lexical activation occurs autonomously rather than through interactive means. Norris et al (2000) presented two main arguments in favor of the modular theory in speech perception. Firstly, they argued that invoking top-down feedback contradicts the principle of simplicity. Secondly, they contended that online top-down feedback does not enhance recognition accuracy at the lexical level, that is to say, if the prelexical process is performing well at the beginning, feedback cannot enhance this performance but could potentially introduce errors.

This viewpoint is supported by evidence showing the absence of contextual effects in the early stages of processing (Twilley & Dixon, 2000). Early studies, such as Swinney's

(1979), support this. Swinney used a method in which listeners, after hearing sentences containing homographs, were showed letter strings related to either the contextually appropriate or inappropriate meanings of the homographs. In Swinney's Experiment 1, listeners listened to a sentence with the ambiguous word *bugs*, which is a homograph that can indicate either insects or listening devices. Listeners heard sentences in both "no context" and "biasing context" conditions, such as *...The man was not surprised when he found several bugs in the corner of the room* (no context) and *...The man was not surprised when he found several spiders, roaches, and other bugs, in the corner of his room* (biasing, *bugs* for insects). Three words, *ANT* (contextually related), *SPY* (contextually inappropriate), and *SEW* (unrelated), were showed on the screen immediately after the word *bugs*. Listeners were asked to quickly decide if the string was a real word or a non-word by pressing a button. Swinney (1979) expected that listeners would likely respond faster to *ANT* (contextually related, "insect") and *SPY* (contextually inappropriate, "listening device") compared to *SEW* (unrelated). The results showed that listeners did respond to *ANT* and *SPY* with similar speed, faster than to the unrelated word *SEW*. This suggests that listeners were faster to recognize words related to both meanings of the homograph, meaning that all meanings are accessed in listeners' minds upon hearing the homograph, regardless of the context ("no context" or "biasing"). In Experiment 2, Swinney used similar materials, but with a key difference: the three visual words were presented on the screen three syllables after the ambiguous word rather than immediately following it (as in Experiment 1). The results showed that when the visual words were presented three syllables later, listeners recognized the contextually related word faster than the contextually inappropriate word. This indicated that the inappropriate meaning was suppressed shortly after the initial access. Therefore, Swinney's experiments demonstrated that while both meanings of an ambiguous word are accessed initially, the contextually irrelevant meanings are suppressed relatively quickly, in this case within three syllables after the ambiguous word. This supports the viewpoint that contextual effects in the early stages of processing could be absent and start to emerge after three syllables (D. Swinney, 1979).

Further, Bowers and Davis (Bowers & Davis, 2004) also illustrated that if top-down feed-

back were present, it could prevent a listener from noticing a mispronunciation in a word, such like substituting the phoneme */d/* for */l/* in the word *elephant*. However, in reality, listeners do typically notice such mispronunciations. This indicates that the initial stages of lexical processing likely operate autonomously, without the influence of top-down feedback, supporting the argument that top-down influences are more relevant in post-access stages of processing.

Therefore, a representative example of a model employing this autonomous approach is the Cohort Model (Heald & Nusbaum, 2014; Marslen-Wilson, 1987). Cohort model also presupposes fixed units in a hierarchical way of activation, and also incorporates bottom-up activation and top-down filtering. The Cohort model posits that speech perception undergoes three phases. In the access phase, auditory lexical retrieval begins with fixed units, such as one or two phonemes (Weber & Scharenborg, 2012). When the phoneme reaches the listener's ear, approximately 150-200ms later, words starting with that specific phoneme or phonemes are all activated in the listener's mental lexicon. Words that are activated but are not the target are referred to as "competitors." In the selection phase, words inconsistent with the incoming segments leave the cohort. The last phase in the Cohort model is integration, where top-down context, as high-level information, constrains the choice of words, and words matching the syntactic and semantic context are ultimately selected. However, the Cohort model emphasizes a bottom-up processing approach where the initial stages of word recognition are influenced solely by the acoustic-phonetic input, without immediate influence from higher-level linguistic or semantic context. Heald and Nusbaum (2014) concluded that the Cohort model treats phonetic features as passively derived and not influenced from expectations.

Moreover, Fuzzy Logical Model of Perception (FLMP) is also founded on the principle that speech perception involves the integration of independent information sources (Massaro, 1989; Massaro & Chen, 2008) .Unlike the Cohort model, which assumes phoneme-sized perceptual units, FLMP maintains that speech perception starts with syllable-sized units. In this model, the characteristics of syllables from the auditory input are aligned with those of corresponding prototypes. These features are then combined into a unified

feature integration. Subsequently, the prototype corresponding to the syllable is activated and identified based on the degree of similarity between the integrated input features and the prototype itself. FLMP does not account for top-down influences from the word level to the phoneme or syllable level, underscoring its focus on the combination of independent information sources without anticipating the impact of top-down feedback on phoneme sensitivity.

Thus, traditional models assume that fixed units are activated hierarchically and acknowledge the interplay between bottom-up and top-down influences. However, as noted above, these models differ in their agreement on the timing of top-down integration during speech perception. Some, such as TRACE, advocate for an early integration, where top-down processes influence the initial stages of speech processing, thereby supporting a non-modular view of cognitive architecture. Conversely, other theories maintain that top-down influences are more relevant during later stages, aligning with a modular approach. This divergence underscores the ongoing debate within cognitive science about the dynamics of perceptual and cognitive systems, and highlights the complexity of speech perception as an interplay of multiple cognitive processes. Therefore, the goal of this thesis for speech unit perception is to examine how flexible that interplay between bottom-up and top-down information among both native and non-native listeners.

2.3.2.2 Models with adaptive resonance

As discussed in the previous Section 2.1 *Speech perceptual units*, previous studies have showed that listeners demonstrate inconsistent results in perceiving fundamental units. Listeners reacted to phonemes, syllables, words, and even phrases containing two or three words. Therefore, the typical candidates for the fundamental speech unit are phoneme and syllable, while other units are still under debate. As mentioned in Section 2.3.1 *Two key issues: activation and representation*, McNeill and Lindig (1973) posited that the "perceptually real" should be what people pay attention to. In normal language use, the focus of attention is the meaning of an utterance. Therefore, unlike models that predict a fixed hierarchical unit, other models focus on predicting resonant dynamics and challenge the notion of a "fixed unit" being activated during speech perception.

Models such as the Ideal Adapter Framework (Kleinschmidt & Jaeger, 2015), MINERVA 2 (Hintzman, 1984, 1986, 1988) and Grossberg's Adaptive Resonance Theory (ART) (1987, 1999, 1976a, 1976b, 1980, 2003, 2013), applied by Goldinger and Azuma in speech perception (Goldinger & Azuma, 2003), do not rely on a fixed hierarchical mechanism. For example, the Ideal Adapter Framework posits that the perceptual unit can include phonetic categories and is influenced by contextual and situational factors, rather than being tied to specific linguistic units like phonemes or syllables alone. MINERVA 2 and ART have a common foundation: both rely on resonant dynamics that flow between working memory and long-term memory (Goldinger & Azuma, 2003). Therefore, MINERVA 2 is similar to ART in that it does not propose fixed units to be activated hierarchically. However, ART is more representative than MINERVA 2 in emphasizing adaptive resonance as an alternative to activating fixed units. Goldinger and Azuma (2003) commented that MINERVA 2 is "limited" relative to ART and that ART is a better theory for advancing beyond MINERVA 2 in arousing adaptive resonance to draw listeners' attention and activate episodic memory (both ART and MINERVA 2 are exemplar theories). As briefly introduced in *Section 2.3.1 Divergence in speech perceptual models and two key issues: activation and representation*, ART proposes *Adaptive Resonance*, or *Resonant Dynamic*, which is an emergent product resulting from the interaction between bottom-up (input signals in working memory) and top-down information (prior knowledge stored in long-term memory).

ART does not deny the existence of perceptual units but incorporates concepts such as phonemes, syllables, and words. It emphasizes that the unit perceived as the perceptual unit depends on which one draws the listener's attention during the interaction between bottom-up and top-down information. Thus, ART is fundamentally interactive not only in information flow but in the very nature of the unit of perception. There are two aspects of adaptive resonance as a key concept in ART for speech perception. First, ART posits that the size and nature of these units are not predetermined but emerge from the interaction of bottom-up and top-down processes (Goldinger & Azuma, 2003). Second, this resonant dynamic is influenced by listeners' attention. Listeners' attention is drawn when bottom-up and top-down knowledge sources achieve resonance, creating a conscious experience.

In other words, ART predicts the perceptual unit, but it is determined by which unit gains the listeners' attention.

In this thesis, the fundamental rationale for the monitoring experiment is from the ART framework (Goldinger & Azuma, 2003). As illustrated before, this framework posits that human listeners' attention can be drawn and directed by either bottom-up or top-down information, or both, to primary perceptual units of different sizes, such as phonemes or syllables, which are not fixed. In other words, both phonemes and syllables can serve as fundamental perceptual units, but their perception can be influenced by the bias of information flow. Typically, syllables, as larger units, can mask phonemes as the primary perceptual unit for listeners. However, phonemes can also dominate syllables when attention is deliberately directed (information flow bias) to them. However, the application of ART in speech perception has been studied among native listeners by Goldinger and Azuma (2003). Therefore, this thesis adopts the ART framework and aims to explore the flexibility of non-native listeners in directing attention between phonemes and syllables, as well as the differences in flexibility between native and non-native listeners. It also seeks to determine the extent to which bottom-up and top-down information can influence the perception of these units among both native and non-native listeners.

2.3.2.2.1 ART information flow ART was first introduced by Grossberg (1999, 1976a, 1976b, 1980, 2003, 1997) as an unsupervised network and a cognitive and neural model of the brain autonomously learning. Later Goldinger and Azuma (2003) illustrated ART's framework as applied in speech perception. ART emphasizes the importance of this interaction, positing that perception occurs when bottom-up and top-down knowledge sources bind into a stable state. Bottom-up information might include the acoustic features of speech, such as the quality of suprasegmental cues or the ambiguity of different phonemes. Top-down information, on the other hand, encompasses the listener's pre-existing knowledge, expectations, and contextual cues that influence how incoming sensory information is interpreted. The bottom-up information, and top-down information are the two factors that cause the resonant brain states and draw attention to the perceptual unit (Grossberg, 1980, 2013; Grossberg & Kazerounian, 2016). However, the weights

of these two parts of information do not need to be equal, which means that the strong bottom-up information can arouse the resonance with the minimal top-down matching, and the distorted or unclear bottom-up signals can be recovered by the support from the top-down prior knowledge.

As illustrated in *Section 2.2 Masking Effect and Reversed Masking Effect as a Phenomenon*, the reversed masking effect was achieved by Goldinger and Azuma (2003) by manipulating both bottom-up (e.g. lengthening fricatives and accentuating stop release to achieve phoneme-biased stimuli) and top-down information (e.g. informing listeners of the primacy of phonemes or syllables in speech perception) to direct listeners' attention to the phoneme-level unit, eventually achieving the reversed masking effect—listeners reacted to phoneme targets faster than to syllable targets. Thus, ART's framework suggests that the cognitive system's ability to integrate bottom-up and top-down information effectively through resonant dynamics is central to understanding speech perception. Therefore, the fundamental unit of speech perception is not fixed in size but rather selected based on cognitive resonance directing attention to it.

2.3.2.2.2 ART working principles The basic ART network architecture consists of four mechanisms: (1) a comparison field, (2) a recognition field, (3) a vigilance parameter, and (4) a reset module. ART is designed with two layers to process the input signal: the comparison field (F1 layer) and the recognition field (F2 layer). The F1 layer comprises input neurons, each represented as a single node. In speech perception, these nodes can be interpreted as the acoustic-phonetic features of the input signals. Each input is described as a vector containing various features as nodes. The F2 layer also consists of nodes, which are products of prior learning stored in memory.

Goldinger and Azuma (2003) interpreted the vector in the F1 layer as "feature clusters" and the nodes in the F2 layer as "list chunks," which include all the possible prior-learned products such as phonemes, syllables, and words. The F1 layer is designed to compare the input pattern with the expectation, while the F2 layer recognizes the best-matched expectation or the best-matched product of prior learning. Each node in the F2 layer can be in one of three states: (1) active, meaning the input best matches this node; (2) inactive,

meaning the input node does not best match the F2 node but is still available to participate in competition; and (3) inhibited, meaning the input node does not match the F2 node and is prevented from participating in further competition.

In cognitive process during speech perception, the vigilance parameter is tested. The first direction is from the bottom to the top, from the F1 layer to the F2 layer. In this connection, the input information is forwarded via the bottom-up weights, and the node in the F2 layer that best matches the input will fire. This means the input is greater than the vigilance parameter ($0 < p \leq 1$), allowing it to match with the node in the F2 layer. The other direction of connection is from the top to the bottom via the top-down weights. If the node is close enough to the input template, it fires as an expectation in the F2 layer, achieving successful categorization and resonance. However, if the input does not exceed the vigilance parameter, the node will not be activated. The cooperation between these two layers and the competition among the nodes of the F2 layer smooth out small mismatches; only large mismatches inhibit the achievement of resonance. If a node is rejected, the system will reset by inhibiting and removing the current expectation. A new competition will then be initiated in the F2 layer. Therefore, resonance is achieved through the process of matching and adapting the most accurate node.

Therefore, in ART, speech perception processing begins when the featural input (in the sense of phonological features) activates the feature cluster (F1 layer), and then the cluster activates the list chunks (F2 layer). Once the items activate list chunks (e.g. phonemes, syllables, words), a feedback cycle begins. If the items receive sufficient top-down confirmation, they will continue to send activation upward. ART predicts that unitization is flexible and the process is opportunistic. The most predictive units will dominate the behavior.

2.3.3 Representation of units

2.3.3.1 Prototype-based representation

Prototype theory in speech perception posits that listeners abstract general, prototypical phonological representations from the detailed specifics of speech during their expe-

riences. These generalized representations are then stored in the cognitive repertoire, serving as reference points against which new speech inputs are compared (K. Johnson, 1997). The framework and working principles of models illustrated in the previous *Section 2.3.2.1 Models with fixed hierarchical mechanism*, such as the TRACE model (McClelland & Elman, 1986), Shortlist (Norris, 1994), Cohort (Marslen-Wilson, 1987), and Fuzzy Logical Model of Perception (FLMP) (Marslen-Wilson, 1987; Massaro, 1989; Oden & Massaro, 1978), predict abstract phonological information as described in prototype theory. These models treat lexical access as a direct mapping from low-level information about the speech signal simultaneously onto a distributed abstract representation of both the form and the meaning of words (Gaskell & Marslen-Wilson, 1997). For example, central to the FLMP are the summary descriptions of the perceptual units of the languages. These summary descriptions are the prototypes and contain the features.

Over time, as listeners accumulate experiences with various exemplars, they may abstract common features from these detailed memories, forming prototypes that capture the essence of a category. Consequently, when encountering new speech sounds, listeners use these prototypes to categorize the sounds based on their similarity to the stored prototypes, facilitating the recognition and understanding of speech across different contexts and speakers. This prototype perspective is supported by numerous studies highlighting the critical role of abstract representations in speech perception (Cutler et al., 2010; Cutler, 2008), suggesting that abstract phonological knowledge facilitates swift adaptation to variability in speech, particularly when encountering different speakers.

Many studies also argued that the abstract representations play a necessary role in speech perception (Cutler et al., 2010; Cutler, 2008). Abstract phonological knowledge provides the evidence of the rapid adaption to speech variability where different talker is at issue. The abstract framework is what enables speakers to recognize and produce speech across various contexts, speakers, and modalities, despite the considerable variability in acoustic signals. Cutler (2017) highlighted the evidence and importance of that the phonological representations are stored in abstract way through five distinct perspectives.

The first evidence is that the L2 speech perceptual persistent difficulties are caused by the L2 phonemic distinctions, highlighting the abstract nature of phonological representations. The challenge for L2 learners lies in modifying this framework to include new phonemic categories that are absent in their L1, a process that involves re-configuring abstract phonological representations to accurately reflect the phonemic inventory of L2. The fact that learners can know, in the abstract, that certain L2 phonemes should be distinct (as in the case of phoneme */l/* in the word *light* and */w/* in *write*) but still fail to perceive and differentiate these sounds in spoken language points to the abstract nature of phonological representations (Broersma & Cutler, 2008; Cutler et al., 2006; Weber & Cutler, 2004). This scenario suggests that understanding and recognizing phonemic distinctions involve more than just learning new sounds or memorizing rules; it requires integrating these sounds into an existing framework of phonological knowledge that is shaped by the learner's L1.

Secondly, during speech perception, listeners swiftly adjust and adapt to new speakers by adjusting phoneme category boundaries, a process that generalizes to previously unheard words and phonetic contexts from the new speaker (McQueen et al., 2006). This adjustment and adaptation demonstrate the necessity of abstract phonological categories: the phonemic categories are at an abstract level, rather than relying solely on detailed traces of specific lexical experiences. Therefore, abstract phonological representations provide a more adaptable and generalized framework for understanding and categorizing speech sounds, essential for effective communication and language comprehension across varying speakers and contexts. Without abstract phonological representations, the episodic traces of lexical experience, which are less flexible and cannot easily generalize across different speaking varieties, cannot deal with this generalisation ability and result (Cutler et al., 2010). Thus, abstract phonological representations enables cognition to go beyond the specifics of individual experiences to categorize the sounds and comprehend the speech.

Thirdly, the consistency in recognizing words across auditory and visual domains, without being influenced by speaker familiarity, underscores the abstract nature of phonolog-

ical representations. These findings highlight that such representations are robust and generalized, not limited by the detailed experiences or the visual identity of speakers. Phonological representations in the lexicon are accessible across both auditory and visual domains, wherein auditory and visual processing reveal speaker-specific information that is not transmitted through these modalities at the lexical level. Abstract representations assist listeners in capturing this information, uninhibited by familiarity with the speaker. Moreover, native listeners exhibit proficiency in recognizing speech from a new speaker added to a familiar group, an ability attributed to phonological familiarity derived from abstract mental lexicon representations (E. K. Johnson et al., 2011).

Finally, children who are adopted into another country and lose all conscious knowledge of their original native language, yet become native speakers of a new language, demonstrate an "accelerated trajectory" in learning the phonological structures of their original native language. This mastery in perception is later transferred, suggesting that such performance benefits from abstract phonological representations, which are compiled for later stages of language acquisition, even before the age of six months (Choi et al., 2017). Thus, Cutler (2017) posits that abstract phonological knowledge is crucial for language acquisition and adaptation.

The prototype theory is based on the recognition of abstract, idealized, context-free symbolic representations. This approach encourages research designed to identify simple first-order acoustic invariants and often overlooks the issues of acoustic-phonetic and indexical variability. Consequently, the prototype, which is the conventional "abstractionist" (or "analytic") approach to speech perception, is criticized for treating variability as a source of "noise" in the speech signal that needs to be eliminated to uncover the "hidden" abstract underlying symbolic linguistic message (Elman & McClelland, 1986). Therefore, exemplar-based lexical representation offers an alternative to the prototype view.

2.3.3.2 Exemplar-based representation

Different from the prototype theory where the prototype or abstract phonological representations are stored and created in cognitive phonological repertoire, exemplar models

offer the alternative to prototype models in the representation of linguistic units. Support for exemplar models comes from evidence that human perceptual system encodes and retains very fine episodic details of speech. Goldinger (1998) illustrated that mental lexicon could be a collection of detailed memory traces, rather than abstract units, and spoken word recognition utilizes highly detailed information in the speech signal, as well as the episodic context, which are stored by the listener, becoming part of a rich and detailed memory representation of speech in long-term memory. Listeners encode and store specific instances of events and their associated episodic contexts, rather than idealized prototypes. Exemplar-based models allow listeners to adapt to and recognize speech from a wide variety of speakers by referencing a rich tapestry of stored examples, each associated with particular speakers. The flexibility of this model lies in its ability to dynamically adjust to new or unfamiliar speakers by activating the most relevant stored instances for comparison.

Models such as ART (Grossberg, 1976a, 1976b, 1980), MINERVA 2 (Hintzman, 1984, 1986, 1988), generalized context model (GCM) (Nosofsky, 1986, 1988, 1991), XMOD (K. Johnson, 2005) and Ideal Adapter Framework (Kleinschmidt & Jaeger, 2015) can commonly describe the exemplar theory.

ART is detailed in the preceding Section 2.3.2.2 and MINERVA 2 is in next Section 2.4 *Working principle of MINERVA 2*. Goldinger and Azuma (2003) viewed the ART as a good improvement of MINERVA 2. MINERVA 2 is similar to ART, predicting that only episodic traces are stored in memory and sharing the common ground with ART. For every known word in MINERVA 2, there is a vast collection of partially redundant traces residing in memory. These traces store the "conceptual," "perceptual," and "contextual" aspects of the events at the time of their initial encoding.

GCM was introduced by Nosofsky (1992, 1986, 1988, 1991, 2000) which is an exemplar model simulating how a new stimulus is classified to a category based on its similarity to the existing members in the category. Therefore, the main feature of GCM is the principle of psychological similarity (Nosofsky, 2011; Shepard, 1987). Applied to speech

perception, GCM assumes that the categories correspond to words, and within each word category, there are word exemplars. Thus, each stimulus can be represented by multiple psychological features which can be considered as multiple psychological dimensions.

The two important concepts that GCM adopted are *multidimensional scaling* (MDS) and *attention weight*. MDS is a tool used by GCM to select perceptual dimensions. Then the stimuli can be represented by their values along the multiple underlying perceptual dimensions. When GCM is applied to speech perception, the multiple dimensions can be the auditory representations, that is, each dimension can be considered as each phonetic property. For example, the formant values F0, F1, F2, F3 and duration can be used as the dimensions of a vowel exemplar psychological space. Similarly, the phonetic properties like the place of articulation, manner of articulation, voice onset time, and duration can also be used as the dimensions of consonant exemplars in a psychological space (K. Johnson, 2005).

The attention weight is used to systematically modify the structure of psychological space during the stimulus categorization process. It can shrink or expand the perceptual space along each auditory dimension (K. Johnson, 1997; Nosofsky, 1992; Nosofsky, 1986) so as to optimize the listeners' classification performance. Thus, the attention weight indicates which auditory properties or dimensions are more valuable in the categorization process, in other words, the categorization process is sensitive to which dimensions are used. Large values of attention weight on the particular dimensions serve to stretch the psychological space along these dimensions. This suggests these dimensions are highly utilized and relevant to the categorization process. On the contrary, small values of attention weight on the particular dimensions serve to shrink psychological space along these dimensions. This suggests these dimensions are not utilized and irrelevant to the categorization process. This stretching and shrinking of the psychological space influences the possibility of a new stimulus can be classified into its category, which further influences the degree of similarity between the input stimulus and the existing exemplar(K. Johnson, 1997; Nosofsky, 1992; Nosofsky, 2011; Vanpaemel & Lee, 2012). For example, the dimension of F1 will gain more attention weight when listeners categorize a set of vowel

stimuli.

XMOD is another exemplar model, proposed by Johnson (K. Johnson, 1997), which has different working principles to GCM. It assumes that the incoming speech signals are transformed into a sequence of spectra indicating multiple phonetic properties. However, XMOD shares some common ground with the GCM. First, the activation of exemplar is based on the degree of similarity to the input stimuli. Second, there are multiple dimensions each considered as different phonetic properties when the model is applied to speech perception.

The Ideal Adapter Framework (IAF) (Kleinschmidt & Jaeger, 2015) is a Bayesian belief updating model. It predicts a *probabilistic process* where listeners use both prior knowledge and new evidence to update their understanding of the current situation. *Probabilistic* in IAF refers to the likelihood or probability of an event or belief occurring. Thus, the probabilistic process means the process that listeners calculates the likelihood or probability that the incoming speech signal matches their prior expectations. If the incoming signal is very likely to match the expected patterns, the belief remains stable. However, if the new signal does not match the prior belief, the listener adjusts their belief based on the probability of the new input. This adjustment is probabilistic, indicating that the listener does not completely discard their prior beliefs but rather refines them based on the relative likelihood of new patterns. Therefore, IAF maintains a balance between stability, which means retaining core knowledge, and flexibility, which indicates adapting to new patterns. Listeners update their prior beliefs incrementally without discarding them entirely.

The IAF can also be viewed as an exemplar model. It assumes that prior beliefs help listeners recognize familiar patterns in acoustic signals based on past experience. Listeners then generalize similar patterns they've encountered before, mapping incoming speech to categories they've previously experienced. When the new signal doesn't align with expected patterns, the listener recognizes this deviation and adapts by updating their belief probabilistically. For example, applying the Ideal Adapter Framework, suppose a listener hears a speaker with a strong accent pronouncing the word "pen" as "pin." Based on the

listener's prior belief, they might be confused by the unexpected pronunciation. In the process of *belief updating*, listeners gather more instances of this pronunciation pattern. In terms of "flexibility," the listener adjusts to a variation in pronunciation that deviates from their prior knowledge or past experience. For "stability," despite this flexibility, the listener retains the ability to correctly interpret the pronunciation of "pen" by other speakers who use the standard pronunciation, maintaining consistent interpretations for familiar pronunciation patterns.

The exemplar theory does not completely deny the existence of prototypes. The concept *exemplar-based speaker normalization* was proposed (K. Johnson, 1997, 2005; Pisoni, D. B., 1997) and is an important theory in the exemplar theory. This theory posits that exemplar-based approach does not preclude the extraction of some level of abstract representation during speech perception. Listeners can still follow abstract general principles about speech sounds from their accumulated experiences. According to this normalization theory, the speaker normalization enables listeners to recognize words spoken by different talkers, despite considerable acoustic variation in phonologically identical utterances across speakers. This process involves the listener's ability to adapt to and categorize speech sounds in a way that accounts for individual speaker characteristics.

Johnson (1997) pointed out that the exemplar theory also involve generalization but this generalization occurs during the retrieval or activation of categories. To be specific, exemplar theory still encompasses generalization, but this generalization occurs during the retrieval or activation of categories, not during their representation or creation. Although exemplar-based models account for this variability by storing detailed memories of how words and sounds are produced by different speakers, some scholars (K. Johnson, 1997; Pisoni, D. B., 1997) take this dynamic generalization view predicting a dynamic process where listeners generalize the detailed, stored instances of speech sounds to adapt to and recognize speech from various speakers.

Some scholars also holds the point that the generalization and the detailed memory traces coexists. For example, Pisoni (1997) acknowledged the existence of abstract and idealized

prototypes within the framework of exemplar-based lexical representation. He (Pisoni, D. B., 1997) posited for *speaker normalization* that abstract and idealized prototypes exist, but argues that rather than arising during simultaneous perceptual analysis, abstractions only emerge from the "computational processes" involved in accessing rich and highly detailed memory representations including intricate details about regional dialects, speaking rates, styles, and other episodic speech characteristics integrating into the long-term knowledge base of listener. This leads to the formation of generalizations not primarily occurring at the point of initial representation of speech sounds in memory, but rather as a dynamic, ongoing process where abstraction and the formation of prototypes result from interacting with and analyzing the detailed contents of memory. Therefore, exemplars and prototypes could coexist, suggesting that human listeners' cognitive systems utilize these two different representations at different times. Initial and simultaneous perceptual analysis relies on exemplars, while later computational processes require prototypes.

2.4 Exemplar models' formulas highlighting cognitive process in native and non-native speech perception

MINERVA 2, as an exemplar model, provides formulas to measure parameters such as similarity distance, activation level, echo content, and echo intensity. As briefly mentioned at the end of *Section 2.3.1*, the monitoring experiment in this thesis seeks to provide additional experimental evidence for models of speech perception. It demonstrates the flexibility of human listeners in perceiving speech units, as well as the extent to which both bottom-up and top-down information influence their perception. In this study, I employed formulas from MINERVA 2 to highlight differences between native and non-native speech perception and to aid in developing hypotheses for the monitoring experiments. Specifically, the formulas from MINERVA 2 are used to describe the cognitive processes involved in native and non-native speech perception.

2.4.1 Working principles and mathematics of MINERVA 2

2.4.1.1 Working principles of MINERVA 2

MINERVA 2, introduced by Hintzman (1984, 1986, 1988), is a computational model designed to describe various cognitive phenomena in human memory. MINERVA 2 predicts that only episodic traces are stored in memory. For every known word in MINERVA 2, there exists a vast collection of traces in memory, which encode the conceptual, perceptual, and contextual details of their original encoding events. Therefore, MINERVA 2 is described as a multiple-trace memory model.

The fundamental concepts in MINERVA 2 are working memory and long-term memory. The working memory receives signal inputs and sends a *probe*, which acts as a retrieval cue, to the long-term memory. The long-term memory contains a large number of episodic memory traces. Once these traces are activated according to the features of the probe sent by the working memory, the long-term memory responds to the short-term memory based on the activation of these traces. This response is called an *echo*.

To be specific, Hintzman (1984) proposed five fundamental working principles for MINERVA 2. First, memory stores traces left by every experience and event. Since MINERVA 2 is a pure exemplar model, it assumes that each experience leaves a unique memory trace. Word representation in speech perception is achieved by activating these stored traces (Goldinger, 1998). Second, the repetition of an experience generates multiple memory traces of a particular word. Thus, for each probe, there are typically multiple memory traces to be activated. Third, the probe activates all memory traces simultaneously, and the echo sent to the working memory is an aggregate of all the activated traces. Fourth, each memory trace is activated according to its similarity to the probe, with the degree of similarity determining the activation level. Fifth, the traces respond simultaneously, and the echo is the sum of the output from these traces.

Based on these principles, MINERVA 2 creates a feature vector to represent each experience. Each experience is presented as a list of feature values classified using three numbers: -1, 0, and +1. The value +1 represents a positive link between the word and

that feature, indicating that this feature facilitates the activation of the word. The value -1 represents an inhibitory link, meaning this feature inhibits the activation of the word. A value of 0 is defined as either an irrelevant or unknown feature for that particular word and feature. The activation process in MINERVA 2 is achieved by comparing the similarity of features between the probe and the trace.

Presentation of the probe in a feature vector is:

(1)

Probe	-1	+1	-1	.....	-1	+1

Presentation of the traces in a feature vector is:

(2)

Trace 1	+1	0	-1	.....	-1	-1
Trace 2	0	+1	+1	.....	-1	0
...	...	...	...	...	...	...
Trace *i*	+1	+1	-1	.....	0	+1

The MINERVA 2 framework involves four key parameters in its mathematical formulation. The first parameter is the *degree of similarity* between the trace and probe. The second parameter is the *activation level* of a trace, calculated as the cube of the degree of similarity. The third parameter is the *intensity of the echo*, which aggregates the activation levels of all traces. The fourth parameter is the *content of the echo*, defined as the sum of the products of the activation level and the feature values of each trace. Hintzman (1984, 1986, 1988) provided explicit formulas for calculating each of these parameters.

MINERVA 2 demonstrates the comparison between incoming input and existing stored traces in memory. This is crucial for explaining how listeners use episode traces from past experiences to process new speech input and handle variations in speech. Thus, the principles of MINERVA 2 allow for hypotheses about how native and non-native listeners differ in their reliance on detailed stored acoustic cues. In the following sections, I detail the implementation of these calculations and illustrate how they provide a mathematical basis for understanding the cognitive processes of speech perception in both native and

non-native contexts, including their differences.

2.4.1.2 Implementation of mathematics in MINERVA 2

The principle of MINERVA 2 is characterized by a structured list of feature values for each memory trace. Memory traces serve as the basic perceptual or retrieval units. I applied MINERVA 2 to describe the cognitive processing of listeners as they encounter speech messages and match these input elements with the stored perceptual units in their cognition.

In this context, MINERVA 2, as a speech perceptual model, can provide explanations for listeners' potential differences observed in native and non-native speech perception. If native listeners respond more slowly to accented fundamental perceptual units—identified here as phonemes and syllables—compared to non-native listeners, who exhibit an advantage when processing accented or distorted speech elements, then, compared to native listeners, non-native listeners are less "sensitive" during speech perception.

Conversely, if non-native listeners exhibit greater sensitivity to and encounter more difficulty with accented perceptual units compared to native listeners, then compared to native listeners, non-native listeners are more "sensitive" during speech perception.

In this section, the formulas from MINERVA 2 used to demonstrate the calculations are: (1) the Euclidean distance between the probe and trace, (2) the activation level, (3) the echo intensity, and (4) the echo content in native and non-native speech perception.

Similarity between probe and traces. To calculate the similarity of each feature between the probe and traces, Hintzman set *P(j)* as the value representing the *j*th feature of a probe and *T(i,j)* as the value representing the corresponding *j*th feature of the trace *i*.

Hintzman set *N* as the total number of features in both the probe and traces, and N_R as the total number of the relevant features for comparison that are nonzero in either *P(j)* or *T(i,j)*. A *relevant feature* here here means that if the feature *j* in either the probe and trace is relevant to the comparison, then either P(j)$\neq$0 or T(i,j)$\neq$0. Thus, N_R=N-Z_I, where Z_I

is the total number of features that are irrelevant for the comparison, which is when both P(j)=0 and T(i,j)=0.

Then, the similarity of the trace *i* to the probe is calculated as:

(3)

$$S(i) = \sum_{j=1}^{N} P(j) * T(i,j)/N_R$$

Calculated from the similarity of trace *i* to the probe, the *activation level* of trace *i* is:

(4)

$$A(i) = S(i)^3$$

Based on the activation level of trace *i*, the *echo intensity* is the sum of the activation level A(i) of all traces corresponding to a particular input word. Hintzman marked the traces from 1 to *m*, where *m* represents the total number of traces in long-term memory. Thus, the echo intensity (EI) is:

(5)

$$EI = \sum_{i=1}^{m} A(i)$$

The *echo content* is calculated by summing the product of the activation level A(i) and all the features T(i,j) in each trace. Thus, the echo content is:

(6)

$$C(j) = \sum_{i=1}^{m} A(i) * T(i.j)$$

I construct a feature vector of a probe-trace(i) pair with a total of seven features:

(7)

Probe	-1	+1	0	0	-1	-1	+1
Trace i	-1	0	0	0	+1	-1	+1

Then, for example, P(2)=+1, and T(i,2)=0. The total number N=7, and Z_I=2. The third and fourth features have values of 0 for both probe and trace, thus, N_R=N-Z_I=7-2=5. S(i) in this example is calculated as:

(8)

$$S(i) = [(-1*-1)+(1*0)+(0*0)+(0*0)+(-1*1)+(-1*-1)+(1*1)]/5 = 0.4$$

Then the activation level of trace i can be calculated as:

(9)

$$A(i) = 0.4^3$$

$$= 0.064$$

Therefore, the values of echo intensity and echo content have a positive correlation to the activation level A(i).

One potential result is that native listeners showed more substantial changes in the masking effect, with the identification of phonemes and syllables being more easily affected by the manipulation of bottom-up and top-down information. In this scenario, non-native listeners are less “sensitive” in perceiving speech unit than native listeners.

Regarding the unit representation in MINERVA 2, this could mean that the feature vector of the accented input speech unit contradicts the feature vector of the traces. There are more inhibitory features (marked as *(-1)*) in the feature vector of the input accented speech units corresponding to the original positive features (marked as *(1)*) of the non-accented speech unit in traces.

For example, suppose the original non-accented speech unit in the memory trace is an voiced alveolar plosive /d/, and the input accented speech unit is a retroflex plosive /ɖ/.

For both native and non-native listeners, I set the /d/ and /ɖ/ to differ in two features in the vector. If native listeners have more detailed representation. This means there should be more positive features defined in the memory trace. Therefore, the input accented phoneme and the non-accented phoneme in the trace are represented by the feature vector for native listeners as follows:

(10)

Trace /d/	+1	+1	+1	0	0	0	+1	+1
Input /ɖ/	+1	+1	+1	0	0	0	*-1	*-1

Then, in native speech perception, the degree of similarity between the input and the trace is calculated as follows:

(11)

$$S(i) = \sum_{j=1}^{N} P(j) * T(i,j)/N_R$$

$$= (1*1) + (1*1) + (1*1) + (0*0) + (0*0) + (0*0) + (1*-1) + (1*-1)/5$$

$$= 0.2$$

The degree of similarity between the accented input /ɖ/ and the non-accented trace /d/ is 0.2.

In non-native speech perception, non-native listeners have less detailed representations of phonemes, resulting in more features marked as 0 in the trace compared to native speech perception. In native speech perception, three features are marked as 0 (see Table 10), whereas in non-native speech perception, five features are marked as 0 (see Table 12). This indicates a lack of detailed definition for these features among non-native listeners. Thus, the degree of similarity between the input and the trace is calculated as:

(12)

Trace /d/	+1	+1	+1	0	0	0	0	0
Input /d̰/	+1	+1	+1	0	0	0	*-1	*-1

(13)

$$S(i) = \sum_{j=1}^{N} P(j) * T(i,j)/N_R$$

$$= (1*1) + (1*1) + (1*1) + (0*0) + (0*0) + (0*0) + (0*-1) + (0*-1)/5$$

$$= 0.6$$

The degree of similarity between the accented input /ːd/ and the non-accented trace /d/ is 0.6. Comparing these results, native speech perception shows a lower degree of similarity between the input and the trace. Consequently, this leads to a lower activation level, echo intensity, and less echo content.

This difference arises because native listeners have more precisely defined features in their vectors that positively activate the non-accented phoneme. In contrast, the features of the accented phoneme act as inhibitors to these positive features. Therefore, when processing accented speech, the contradictions encountered by native listeners are much more significant than those faced by non-native listeners.

Another potential outcome is that non-native listeners could exhibit greater fluctuations in the masking effect, suggesting that their RTs to phonemes and syllables could be more dramatically influenced by bottom-up or top-down information compared to native listeners. In this scenario, non-native listeners may be more “sensitive” than native listeners. According to the unit representation in MINERVA 2, while native listeners may possess more detailed feature vectors leading to greater contradictions during trace-input matching, non-native listeners with non-high proficiency in L2 could have inaccurately stored features in their vectors. This could result in non-native listeners facing more challenges in matching the input with the trace.

2.5 Summary

As outlined in this chapter, models of speech perception diverge around two key issues: unit activation and representation. Classical models posit a fixed hierarchical unit activation system, where the flow of information can be either interactive or autonomous. In contrast, more dynamic models propose adaptive resonance, where the information flow is typically interactive due to the interplay between top-down and bottom-up information.

Regarding unit representation, exemplar theory and prototype theory differ: exemplar theory argues that lexical representation is based on detailed memory episodes from experiences, while prototype theory posits that it is represented as prototypes. However, it is important to note that exemplar theory does not entirely deny the existence of prototypes; it suggests that generalization occurs during the retrieval and activation of categories, but not in their representation.

Moreover, Adaptive Resonance Theory (ART), as an exemplar theory, predicts human listeners' flexibility in perceiving units, as demonstrated by the masking effect and the reversed masking effect in the experiments conducted by Goldinger and Azuma (2003). Therefore, ART is highlighted in this thesis. ART posits that speech units are not activated in a fixed hierarchical order; rather, the activation of a phoneme or syllable depends on the size of the unit that captures listeners' attention through the interaction of bottom-up and top-down information, ultimately achieving adaptive resonance for that specific unit.

Consequently, this thesis adopts the framework of ART and aims to explore the flexibility demonstrated by human listeners in perceiving speech units. Additionally, MINERVA 2 is applied in this thesis to develop hypotheses about listeners' cognitive processes during native and non-native speech perception.

Chapter 3

Monitoring experiment method: native and non-native listeners' attention shifting across speech units

3.1 Experiment overview

Phoneme and syllable monitoring is defined as a task in which participants are asked to detect and react to target phonemes and syllables that may appear anywhere in the test words (Frauenfelder & Segui, 1989; Mehler et al., 1981). Participants listen for a particular unit, such as the phoneme /t/ or the syllable /taf/, and press a button as quickly as possible when they hear it.

As mentioned in Section 2.3.2 *Models with adaptive resonance*, the key theory underpinning the current phoneme and syllable monitoring experiment (hereafter referred to as the "monitoring experiment") in this thesis is the framework of ART. The fundamental rationale of the experiment is that human listeners' attention can be directed by bottom-up information, top-down information, or both, to achieve adaptive resonance on either a phoneme or a syllable, depending on where the attention is drawn. Thus, the primary perceptual unit is not fixed, and listeners can flexibly perceive the sizes of speech units. Typically, syllables can mask phonemes.

However, when attention is deliberately directed to phonemes, listeners are able to react to phonemes faster than to syllables, as the adaptive resonance draws attention to phonemes, which is a reversed masking effect. This "information flow bias" can be achieved in two

ways, as showed by Goldinger and Azuma (2003): by manipulating bottom-up input (for example by lengthening fricatives and accentuating stop releases); or by manipulating top-down information (for example by informing listeners that phonemes are a fundamental unit in speech perception). Therefore, listeners in native speech perception showed their "flexibility" in detecting different sizes of perceptual unit as the primary speech perceptual unit. However, this flexibility of listeners has only been demonstrated in native speech perception.

In the current monitoring experiment, I use the concept of *perceptual flexibility* (see Section 1.1 in Chapter 1). Recall that perceptual flexibility in speech perception refers to a cognitive ability that allows listeners to detect primary speech perceptual units of varying sizes. This flexibility is demonstrated by listeners shifting attention between phonemes and syllables. When listeners' attention is effectively directed by the interaction of information flow, adaptive resonance emerges and guides their attention to the appropriate unit level. Conversely, if listeners' attention is not directed by this interaction, adaptive resonance does not emerge, and their attention cannot be focused on the intended unit level.

Listeners with "high perceptual flexibility" can actively and flexibly shift their attention between phonemes and syllables to adapt to signal changes, while listeners with "low perceptual flexibility" maintain a fixed focus on a specific speech unit. In this experiment, perceptual flexibility is operationalized through the concept of the masking effect.

The main goal of the current monitoring experiment in this thesis is to explore the flexibility in perceiving speech units. It examines the flexibility of non-native listeners and differences between native and non-native listeners in directing their attention between phonemes and syllables when the information flow is manipulated. Specifically, the experiment tests listeners' ability to achieve adaptive resonance at the phoneme level when acoustic cues (bottom-up information) and prior knowledge (top-down information) direct their attention to the phoneme to reverse the masking effect.

Therefore, the broad question that this monitoring experiment asks is whether non-native

listeners exhibit the same degree of perceptual flexibility as native listeners. Since fluctuations in the masking effect indicate perceptual flexibility, in other words, the question becomes whether non-native listeners show similar fluctuations in the masking effect as native listeners when bottom-up, top-down, or both types of information are manipulated.

This monitoring experiment is based on the rationale from Goldinger and Azuma's study (2003) and replicates the monitoring experiment in it, though it is not identical. The specific ways in which information flow is manipulated differ from those in prior research. As previously summarized, Goldinger and Azuma's study "boosted" phoneme detection by bottom-up and top-down information, leading to the expectation that listeners would respond faster to phonemes than to syllables (a reversed masking effect). They demonstrated that adaptive resonance (attention) can be flexibly achieved at both the phoneme and syllable levels. The current monitoring experiment aims to investigate whether this flexible attentional allocation across phonemes and syllables can also be achieved by non-native listeners. The experiment creates a challenging context for non-native listeners, utilizing accented sounds as bottom-up information. Unlike Goldinger and Azuma's study, the current experiment does not expect a reversed masking effect because the bottom-up and top-down information given at phoneme-level units does not "boost" phoneme detection. The bottom-up information in this experiment is an artificial accent; therefore, the current study expects that listeners' perceptual flexibility will be reflected in fluctuations in the size of the masking effect, as indicated by longer reaction times for phonemes compared to syllables. This could suggest that listeners experience greater processing difficulty and allocate more time and attention to phoneme-level units and their processing.

In this experiment, when bottom-up information is manipulated, an artificial accent is applied to selectively direct attention to phonemes. This artificial accent involves certain phonemes being produced as variants that do not originally exist in English or Mandarin, such as retroflex, implosive, ejective, glottal, voiceless fricative, and palatal sounds. Thus, the artificial accent used in this experiment distorts the original pronunciation of the phoneme. This approach differs from Goldinger and Azuma's research, where bottom-up information was manipulated in the stimuli by instructing speakers that either phonemes

or syllables were the fundamental unit. This led speakers to produce stimuli with strong acoustic cues, such as lengthened fricatives and pronounced stop releases, making the phonemes more salient without distorting them.

When top-down information is manipulated, listeners are informed of the characteristics of the artificial accent. This also differs from Goldinger and Azuma's study, where listeners were given a phoneme-biasing expectation, such as the expectation that phoneme is fundamental perceptual unit. In the current monitoring experiment, the top-down information directs listeners' attention to the artificial accent and the units carrying this accent, but does not directly bias the phoneme as the fundamental unit.

English and Mandarin were used as the stimulus languages. In addition to examining the perceptual flexibility of non-native listeners, the current experiment also investigates whether the flexibility is language-specific. Hence, two different languages were utilized for this purpose.

3.1.1 Research questions

The general question this experiment asks is whether, regardless of language stimuli, non-native listeners show the same flexibility in masking effect that native listeners are predicted to show? As showed by the prior study (Goldinger & Azuma, 2003), native listeners' attention can be directed by adaptive resonance, resulting from the integration of bottom-up and top-down information. This allows native listeners to flexibly detect the primary speech perceptual unit of different sizes. Specifically, there are two research questions for the monitoring experiment:

Question 1: Is the masking effect different for native and non-native listeners?

Question 2: Does the masking effect on native and non-native listeners show variability (perceptual flexibility) according to factorially manipulating bottom-up and top-down information? (manipulating bottom-up information; manipulating of top-down information; and manipulating both)

3.1.2 Predictions

Perceptual flexibility is manifested through the fluctuation in the size of the masking effect. The masking effect is essentially the difference between the RTs to phonemes and syllables (masking effect is calculated as RTs to phonemes minus RTs to syllables; see Chapter 2, Section 2.2 *Masking Effect as a Phenomenon*, and Chapter 4, Section 4.1.2 *Calculation of Masking Effect*). A significant fluctuation in the masking effect indicates that RTs to phonemes and syllables vary, suggesting that one of them could be more sensitive to the manipulation of information and potentially receive more attention from listeners.

Since the artificial accent, as bottom-up information, is applied to phoneme, RTs to phonemes and syllables carrying the artificial-accented phoneme are expected to increase for both native and non-native listeners. Prior knowledge, as top-down information about the artificial accent, gives listeners an expectancy of the artificial accent, which is expected to mitigate the challenges created by the artificial accent. Thus, RTs to phonemes and syllables are expected to decrease with prior knowledge applied.

Prediction 1: The masking effect is predicted to differ between native and non-native listeners. Specifically, the masking effect is expected to vary more significantly for native listeners than for non-native listeners.

For native listeners, higher perceptual flexibility is predicted, indicated by more substantial changes in the size of the masking effect compared to non-native listeners. Since the artificial accent and prior knowledge of the artificial accent are applied to the phoneme, native listeners' attention is anticipated to be flexibly directed to the phoneme. The increase in RTs for phonemes should be more significant than the increase in RTs for syllables. Therefore, an increase in the masking effect is anticipated for native listeners, who are predicted to be "sensitive" during speech unit perception under than manipulation of top-down and bottom-up information.

For non-native listeners, reduced perceptual flexibility is predicted. Non-native listeners

are not expected to be as sensitive to the artificial accent and prior knowledge as native listeners, making it harder for their attention to be flexibly directed to the phoneme. Thus, an insignificant variation in the masking effect is anticipated for non-native listeners. Therefore, compared to native listeners, non-native listeners are predicted to be less "sensitive."

Prediction 2: Simultaneously applying artificial accent (bottom-up information) and prior knowledge of the artificial accent (top-down information) should have a stronger influence on directing listeners' attention to the phoneme compared to only applying one of these.

3.1.3 Variables

This experiment uses a 2*2*2*2 factorial design with 4 independent variables (IVs), resulting in 16 conditions. RTs to phoneme and syllable targets are recorded to calculate the Masking effect as the dependent variables. The 4 independent variables are as follows.

IVs 1: Target types (Phoneme versus Syllable): Two types of targets were presented for participants to detect and react to: the word-initial phoneme (e.g. /d/) and the word-initial syllable (e.g. /dæg/). The masking effect is labeled by referring to the difference between the RTs for phoneme and its corresponding counterpart syllable (e.g. phoneme /d/ matches with syllable /den/, masking effect is labeled as /d/), as detailed in Tables 4.1 and 4.2 in Section 4.1.2 *Calculation of Masking Effect*. Therefore, the phoneme and syllable targets are matched according to their initial sound. The target type is both a within-subject and a within-item variable: each listener in the study was presented with both phoneme and syllable targets, and each target label represents both phoneme and syllable.

IVs 2: Accent (Standard British/Mandarin accent, in Experiment 1 and 2 respectively, versus Artificial accent): Bottom-up information was manipulated through the Accent variable in two conditions. Artificial accent was applied to the word-initial phoneme for both phoneme and syllable targets. During the experiment, each listener was exposed to two speakers: the first speaker produced stimuli in Standard accent, and the second

speaker produced stimuli in an artificial accent.

IVs 3: Prior knowledge (General knowledge versus Accent knowledge):Top-down information was manipulated through the Prior Knowledge variable in two conditions. In the General knowledge condition, participants were informed that there were two speakers and listened to two recorded self-introductions by the speakers using their own voices. These self-introductions included general information about the speakers, such as where they were from and their daily hobbies. The self-introduction audio recordings did not include any of the target sounds used in the experiment. Thus, the participants were familiar with the voices of the two speakers, but they were not informed that the speakers had accents, nor were they given any information about the characteristics of the accents.

In the Accent knowledge condition, listeners were not only given the information listed above but were also informed that the speakers had accents and that the phonemes and syllables they were about to hear were produced with a specific pronunciation. Additionally, participants were given examples of words pronounced with the speakers' accents.

The Prior knowledge is a within-item and between-subject variable: half of the participants were placed in the Accent knowledge condition, and the other half were placed in the General knowledge condition. Each stimulus was presented under both Accent knowledge and General knowledge conditions.

IVs 4: Modality (In lab/Online): This is a between-subject and within-item variable. Participants were divided into two groups, with half completing the experiment online in their own locations, while the other half completed the experiment in the university's phonetic lab. Stimuli were presented both in the lab and online. The purpose of using online groups is to recruit a sufficient number of participants for each language-speaking group.

3.2 Methods

3.2.1 Experiment design

The current monitoring experiments consists of two parts: Experiment 1 and Experiment 2. Experiment 1 was conducted in English instruction and accents, with English pseudo-word stimuli. Experiment 2 was conducted in Mandarin instruction and accents, with Mandarin pseudo-word stimuli. See Appendix A for all English and Mandarin stimuli. The design of the two experiments were exactly the same, only with the stimuli differing.

Listeners were provided with a target sound, such as the phoneme /d/ or the syllable /dæg/, and were asked to listen for that sound in a set of nonsense pseudo-words. If the sound was present in one of the words, they had to press a "yes" button.

There were four lists in total—List A, B, C, and D. These lists were randomly assigned to listeners, with each listener completing only one list. Each list contained 120 pseudo-words, evenly distributed across 20 blocks, with 6 pseudo-words per block. Example of one list of stimuli from Experiment 1 is given in Table 3.1. Stimuli for Experiment 1 and 2 are detailed in next Section 3.2.2 *Materials*.

Blocks 1 to 10 were produced in a standard accent (British for Experiment 1, Mandarin for Experiment 2). Stimuli in blocks 1 to 5 were for phoneme targets. Stimuli in blocks 6 to 10 were for syllable targets. Blocks 11 to 20 were produced in an artificial accent. Stimuli in blocks 11 to 15 were for phoneme targets, and stimuli in blocks 16 to 20 were for syllable targets. Thus, listeners reacted to phoneme targets when they heard stimuli from blocks 1 to 5 and blocks 11 to 15, and to syllable targets when they heard stimuli from blocks 6 to 10 and blocks 16 to 20.

The experiment design was counterbalanced. Stimuli in Lists C and D were the same as those in Lists A and B. For Accent, List A's (and List C's) pseudo-words produced in an artificial accent were produced in a standard accent in List B (and List D), and vice versa, List A's (and List C's) pseudo-words produced in a standard accent were produced in an artificial accent in List B (and List D). Thus, the Accent as bottom-up information

was a within-subject and within-item variable. Each listener reacted to both conditions of standard and artificial accents, and each stimulus was presented to listeners in both the standard and artificial accents.

For Prior knowledge, Lists A and B were given to listeners without any knowledge of the artificial accent, while Lists C and D were given to listeners with prior knowledge of the artificial accent. Thus, Prior knowledge as top-down information, was a between-subject and within-item variable. Each stimulus was presented under both conditions—with and without prior knowledge—and these two conditions were tested by different listeners.

As for the targets and foils, each block contained different numbers of targets and foils to prevent listeners from detecting a pattern in the target numbers. As showed in Table 3.1, the pseudo-words in bold represent the targets. Each participant responded to a total of 120 pseudo-words, with 50 of these being targets.

block1	block2	block3	block4	block5	block6	block7	block8	block9	block10
/d/	/n/	/b/	/k/	/l/	/den/	/ned/	/bɪs/	/kɔm/	/laɹ/
telson	mastet	**basline**	**capnar**	rasto	dagtill	**nedwawn**	**biscri**	**combete**	losken
tengion	**nitnought**	pirthly	**kipsill**	rumcawd	**dentit**	nicknict	balble	curnave	lotler
descate	mitfle	**bilwit**	ganvet	**lantist**	disnawst	nurbike	**bisthawt**	carbeme	**larmest**
tispin	**netleme**	**beshpoy**	**kinchap**	racta	dirchous	neakful	bemple	**compats**	lofpoys
geplo	masto	pedlish	tidloke	rispawt	dabtice	**nedbake**	bensust	**compins**	**larsark**
pilkag	mirtith	mudgewawn	**cusbine**	**louslo**	dadtike	nicknill	**bisfeld**	**compird**	lompon

block11	block12	block13	block14	block15	block16	block17	block18	block19	block20
/ɖʃ/	/ɲ/	/ɓ/	/kˀ/	/ɬ/	/ɖen/	/ɲed/	/ɓɪs/	/kˀɔm/	/ɬaɹ/
tompom	**nethless**	gultish	panco	**lintass**	**dentin**	**nedlert**	backmoot	cursake	lompest
duplic	**nuptar**	dispin	**casman**	rambirt	dagtile	nurtar	**bisfed**	carmeen	**larfo**
dispa	yampus	gungro	purstol	**linty**	**denfle**	**nedbass**	besspoy	**combose**	lubchous
dapmice	yacten	**busthess**	puptish	rimbus	**dentawss**	neeptics	befter	custeen	**larno**
tinlish	**nesstish**	darvo	tambit	**lampin**	dishast	**nedleme**	belden	curbost	**larlish**
dacma	yimbirt	**bilbood**	tanfish	randap	**dentho**	nickdact	**bisred**	karsact	lusquer

Table 3.1: Example list of stimuli from Experiment 1

3.2.2 Materials

3.2.2.1 Experiment 1: English pseudo-word stimuli

Experiment 1 used English stimuli. All English stimuli are pseudo-words, which are generated from the Pseudo-word generator Wuggy (http://crr.ugent.be/programs-data/wuggy). All pseudo-words conform to the phonological patterns of English. All the English pseudo-word stimuli are two-syllable pseudo-words in a structure of CV(C).C(C)V(CC). There are six types in total. The stress is on the first syllable. The examples of English pseudo-word types are given in Table 3.2.

CVC.CVC	CVC.CVCC	CVC.CCV	CVC.CV	CV.CVC	CV.CVCC
e.g. descate /ˈdeskət/	disnawst/ˈdɪsnɔːst/	gungro /ˈgʌŋgɹəʊ/	dispa /ˈdɪspə/	cursake/ˈkɜːseɪk/	curbost /ˈkɜːbɔːst/

Table 3.2: English pseudo-word types and examples

An artificial accent was applied to word-initial phonemes for both phoneme and syllable targets. The artificial accent incorporates acoustic features that are not typically found in English. There were five phoneme targets: voiced alveolar plosive /d/, alveolar nasal /n/, voiced bilabial plosive /b/, voiceless velar plosive /k/, and alveolar lateral approximant /l/, which were presented in a standard British accent. When produced in an artificial accent, the phoneme targets above were presented in their accented counterparts: retroflex /ɖ/, palatal /ɲ/, implosive /ɓ/, ejective /k'/, and voiceless alveolar lateral fricative /ɬ/.

The syllable targets produced in a standard British accent were: /den/, /ned/, /bɪs/, /kɔm/, and /laɹ/. When produced with the artificial accent, these syllable targets were changed into /ɖen/, /ɲed/, /ɓɪs/, /k'ɔm/ and /ɬaɹ/. The phoneme and syllable targets for English pseudo-word stimuli are showed in Table 3.3.

Block1-5 (P)	Block6-10 (S)	Block11-15 (AP)	Block16-20 (AS)
/d/	/den/	/ɖ/	/ɖen/
/n/	/ned/	/ɲ/	/ɲed/
/b/	/bɪs/	/ɓ/	/ɓɪs/
/k/	/kɔm/	/k'/	/k'ɔm/
/l/	/laɹ/	/ɬ/	/ɬaɹ/

Table 3.3: Phoneme and syllable targets for English pseudo-word stimuli. The column for phoneme targets in standard British accent is marked as *P*, and column for phoneme targets produced with an artificial accent is marked as *AP*. The column for syllable targets in standard British accent is marked as *S*, and the column for syllable targets produced with an artificial accent is marked as *AS*.

English pseudo-word stimuli from Lists A, B, C, and D are given in Tables 3.4 and 3.5. Targets are given in bold. Blocks 1 to 5 and blocks 11 to 15 are phoneme-target blocks. In these blocks, the initial phoneme of foils and targets differs in either the place or manner of articulation. For example, in block 1, the target is the voiced alveolar plosive /d/ (e.g. /d/ in *descate*), so the initial phonemes of foils are designed to be /t/, /g/, or /p/, which share either the alveolar place or the manner of articulation with /d/. The vowels following the initial phonemes in these pseudo-words are designed to vary in position, ranging from high to low and front to back.

Blocks 6 to 10 and blocks 16 to 20 are syllable-target blocks. In these blocks, foils have identical initial phoneme onsets as the targets but differ in vowel and coda (e.g. /den/ in *dentit* versus /dɪs/ in *disnawst*). The vowels in the initial syllables of foils and targets are in neighboring articulatory positions. For example, in block 6, the first vowel in the target woord *dentit* is the front non-low vowel /e/, so the first vowels in the foils are all non-back and non-low vowels like /i/, /æ/, and /ɛ/.

block1	block2	block3	block4	block5	block6	block7	block8	block9	block10
/d/	/n/	/b/	/k/	/l/	/den/	/ned/	/bɪs/	/kɔm/	/laɪ/
telson	mastet	**basline**	**capnar**	rasto	dagtill	**nedwawn**	**biscri**	**combete**	losken
tengion	**nitnought**	pirthly	**kipsill**	rumcawd	**dentit**	nicknict	balble	curnave	lotler
descate	mitfle	**bilwit**	ganvet	**lantist**	disnawst	nurbike	**bisthawt**	carbeme	**larmest**
tispin	**netleme**	**beshpoy**	**kinchap**	racta	dirchous	neakful	bemple	**compats**	lofpoys
geplo	masto	pedlish	tidloke	rispawt	dabtice	**nedbake**	bensust	**compins**	**larsark**
pilkag	mirtith	mudgewawn	**cusbine**	**louslo**	dadtike	nicknill	**bisfeld**	**compird**	lompon

block11	block12	block13	block14	block15	block16	block17	block18	block19	block20
/ɕ[/	/ɲ/	/ɓ/	/kˀ/	/ɬ/	/ɖen/	/ɲed/	/ɓɪs/	/kˀɔm/	/ɬaɪ/
tompom	**nethless**	gultish	panco	**lintass**	**dentin**	**nedlert**	backmoot	cursake	lompest
duplic	**nuptar**	dispin	**casman**	rambirt	dagtile	nurtar	**bisfed**	carmeen	**larfo**
dispa	yampus	gungro	purstol	**linty**	**denfle**	**nedbass**	besspoy	**combose**	lubchous
dapmice	yacten	**busthess**	puptish	rimbus	**dentawss**	neeptics	befter	custeen	**larno**
tinlish	**nesstish**	darvo	tambit	**lampin**	dishast	**nedleme**	belden	curbost	**larlish**
dacma	yimbirt	**bilbood**	tanfish	randap	**dentho**	nickdact	**bisred**	karsact	lusquer

Table 3.4: English stimuli List A and C

block1	block2	block3	block4	block5	block6	block7	block8	block9	block10
/d/	/n/	/b/	/k/	/l/	/den/	/ned/	/bɪs/	/kɔm/	/laɪ/
tompom	**nethless**	pultish	ganco	**lintass**	**dentin**	**nedlert**	backmoot	cursake	lompest
duplic	**nuptar**	mispin	**casman**	rambirt	dagtile	nurtar	bisfed	**carmeen**	larfo
dispa	mampus	pungro	gurstol	**linty**	**denfle**	**nedbace**	besspoy	combose	lubchous
dapmice	macten	*busthess*	guptish	**rimbus**	**dentoss**	neeptics	befter	custine	**larno**
tinlish	**nesstish**	marvo	tambit	lampin	dishast	**nedleme**	belden	curbost	**larlish**
dacma	mimbirt	**bilbood**	tanfish	randap	**dentho**	nickdact	bisred	karsact	lusquer

block11	block12	block13	block14	block15	block16	block17	block18	block19	block20
/ɖ/	/ɲ/	/ɓ/	/kʼ/	/ɦ/	/ɖen/	/ɲed/	/ɓɪs/	/kʼɔm/	/ɦaɪ/
telson	yastet	basline	capnar	rasto	dagtill	**nedwawn**	**biscri**	**combete**	losken
tengion	**nitnot**	girthly	kipsill	**rumcawd**	**dentit**	nicknict	balble	curnave	lotler
descate	yitfle	girthly	**tanvet**	lantist	disnost	nurbike	**bistawt**	carbeme	**larmest**
tispin	**netleme**	beshpoy	kinchap	*racta*	dirchous	neakful	bemple	**compats**	lofpoys
zeplo	yasto	**gedlish**	pidloke	rispawt	dabtice	**nedbake**	bensust	**compins**	**larsark**
zilkag	yirtith	gudgewawn	cusbine	**louslo**	dadtike	nicknill	*bisteld*	**compird**	lompon

Table 3.5: English stimuli List B and D

3.2.2.2 Experiment 2: Mandarin pseudo-word stimuli

Experiment 2 used Mandarin stimuli. All Mandarin stimuli are two-syllable pseudo-words which are generated from real Mandarin words by modifying the tones of characters and the first phoneme of the first character. For example, *di4 lu3 /ti lu/* (tones are marked in 1, 2, 3 and 4) is derived from the real Mandarin word *bi4 lu2 "fireplace"*. All Mandarin stimuli conform to the phonological patterns of Mandarin. Since one Mandarin character is equal to one syllable, a Mandarin two-character word contains two syllables. In all Mandarin pseudo-word stimuli, the first syllable is in tone 4. The stimuli have a structure of CV(C).CV(C), with four types in total. Examples of Mandarin pseudo-word types are given in Table 3.6.

CVC.CVC	CVC.CV	CV.CVC	CV.CV
e.g. ning4 lang2 /niŋ laŋ/	han4 kao4 /xan kʰau/	hu4 zhang3 /xu ʈ͡saŋ/	di4 lu3 /ti lu/

Table 3.6: Mandarin pseudo-word types and examples

Similar to English stimuli, an artificial accent was applied to word-initial phonemes for both phoneme and syllable targets. The artificial accent incorporates acoustic features that are not typically found in Mandarin. There are 5 phoneme targets produced in standard Mandarin: aspirated voiced alveolar plosive /tʰ/, alveolar nasal /n/, voiceless velar fricative /x/, aspirated velar plosive /kʰ/, and alveolar lateral approximant /l/. With artificial accent, these phoneme targets change into their counterparts as aspirated retroflex plosive /ʈʰ/, palatal nasal /ɲ/, voiceless glottal fricative /h/, velar ejective /k'/, and voiceless alveolar lateral fricative /ɬ/.

The syllable targets produced in standard Mandarin are /tʰan/ (tan4), /niŋ/ (ning4), /xən/ (hen4), /kʰʊŋ/ (kong4), and /lau/ (lao4). When produced with the artificial accent, these syllable targets change into their counterparts as /ʈʰan/ (tan4), /ɲiŋ/ (ning4), /hən/ (hen4), /k'ʊŋ/ (kong4), and /ɬəu/ (lao4). The phoneme and syllable targets for Mandarin pseudo-word stimuli are showed in Table 3.7.

Block1-5 (P)	Block6-10 (S)	Block11-15 (AP)	Block16-20 (AS)
/tʰ/	/tʰan/	/ʈʰ/	/ʈʰan/
/l/	/lau/	/ɬ/	/ɬəu/
/n/	/niŋ/	/ɲ/	/ɲiŋ/
/x/	/xən/	/h/	/hən/
/kʰ/	/kʰʊŋ/	/k'/	/k'ʊŋ/

Table 3.7: Phoneme and syllable targets for Mandarin pseudo-word stimuli. The column for phoneme targets in standard Mandarin accent is marked as *P*, and column for phoneme targets produced with an artificial accent is marked as *AP*. The column for syllable targets in standard Mandarin accent is marked as *S*, and the column for syllable targets produced with an artificial accent is marked as *AS*.

Mandarin pseudo-word stimuli from Lists A, B, C, and D are given in Tables 3.8 and 3.9, with targets indicated in bold. Same as the design of the English stimuli in Experiment 1, blocks 1 to 5 and blocks 11 to 15 are phoneme-target blocks. In these blocks, the initial phonemes of foils and targets differ in either the place or manner of articulation (e.g. /d/ versus /t^h/). The vowels following these initial phonemes are designed to vary in position, ranging from high to low and front to back. Blocks 6 to 10 and blocks 16 to 20 are syllable-target blocks. In these blocks, foils have identical initial phoneme onsets as the targets but differ in the vowel and coda (e.g. /niŋ/ versus /nau/). The vowels in the initial syllable of both foils and targets are in neighboring articulatory positions.

block1	block2	block3	block4	block5	block6	block7	block8	block9	block10
/t^h/	/n/	/x/	/k^h/	/l/	/t^han/	/niŋ/	/xən/	/k^hʊŋ/	/lau/
di4 lu3	mao4 zhu1	**hai4 ji4**	**kuan4 dui4**	ruan4 ji2	ti4 bing1	**ning4 ji2**	**hen4 zong1**	**kong4 tai4**	lie4 lin2
dai4 cai4	**nan4 hui4**	gai4 piao1	**kong4 ji2**	ru4 lun4	**tan4 ju2**	nang4 shi2	hai4 li3	kai4 zhong4	long4 ke4
tao4 kuo4	mang4 bai4	**han4 kao4**	gang4 li4	**lan4 qi2**	tu4 cong4	nao4 li2	**hen4 fu1**	ke4 rong2	**lao4 di4**
dao4 hu1	**ning4 lang2**	**hu4 zhang3**	**ku4 yuan2**	ren4 chi4	tie4 yong3	nai4 zong1	hao4 ni4	**kong4 dai4**	lu4 wen1
deng4 fa4	man4 ding4	gan4 zan1	pai4 hua2	rao4 di1	tian4 quan1	**ning4 li2**	hu4 di4	**kong4 li3**	**lao 4 ku1**
dian4 sai4	man4 xiang3	gou4 jue2	**ke4 tu2**	**liang4 gao1**	teng4 jie2	neng4 hao1	**hen4 ling2**	**kong4 tan2**	liu4 xi2

block11	block12	block13	block14	block15	block16	block17	block18	block19	block20
/t^h/	/ɲ/	/h/	/k'/	/ɬ/	/t^han/	/niŋ/	/hən/	/k'ʊŋ/	/ɬəu/
dai4 qi1	**neng4 xi1**	tao4 bu3	peng4 liang4	**lu4 tao1**	**tan4 zheng3**	**ning4 cheng1**	hua4 pian1	kai4 yan4	luo4 pan4
tang4 feng1	**ni4 zhuang1**	te4 huan3	**ke4 huo3**	ran4 biao3	dui4 pan4	nao4 pan4	**hen4 teng2**	kan4 sheng1	**lao4 lun4**
tai4 chang2	yan4 du4	tan4 wen1	peng4 xing1	**lian4 shu4**	**tan4 qian1**	**ning4 fu4**	hu4 pian1	**kong4 fang2**	lu4 ban4
tu4 zhe2	ya4 dou3	**hai4 liang2**	pou4 zi3	ru4 hao2	**tan4 yun4**	nan4 jiang3	hou4 xiao1	ka4 jia1	**lao4 yan2**
deng4 bei4	**nian4 feng1**	tang4 cai2	pu4 lun2	**liao4 gui1**	duo4 ban4	**ning4 gao1**	hong4 dian4	ku4 fu2	**lao4 hang2**
teng4 de2	yan4 deng1	**hui4 jie2**	pan4 shu3	ruo4 ba3	**tan4 hui3**	na4 pu4	**hen4 shi4**	kuo4 de2	li4 su1

Table 3.8: Mandarin stimuli List A and C

block1	block2	block3	block4	block5	block6	block7	block8	block9	block10
/t^h/	/n/	/x/	/k^h/	/l/	/t^han/	/niŋ/	/xən/	/k^hʊŋ/	/lau/
dai4 qi1	**neng4 xi1**	gao4 bu3	peng4 liang4	**lu4 tao1**	**tan4 zheng3**	**ning4 cheng1**	hua4 pian1	kai4 yan4	luo4 pan4
tang4 feng1	**ni4 zhuang1**	ge4 huan3	**ke4 huo3**	ran4 biao3	dui4 pan4	nao4 pan4	**hen4 teng2**	kan4 sheng1	**lao4 lun4**
tai4 chang2	mang4 du4	gan4 wen1	peng4 xing1	**lian4 shu4**	**tan4 qian1**	**ning4 fu4**	hu4 pian1	**kong4 fang2**	lu4 ban4
tu4 zhe2	ma4 dou3	**hai4 liang2**	pou4 zi3	ru4 hao2	**tan4 yun4**	nan4 jiang3	hou4 xiao1	ka4 jia1	**lao4 yan2**
deng4 bei4	**nian4 feng1**	gang4 cai2	pu4 lun2	**liao4 gui1**	duo4 ban4	**ning4 gao1**	hong4 dian4	ku4 fu2	**lao4 hang2**
vteng4 de2	man4 deng1	**hui4 jie2**	pan4 shu3	ruo4 ba3	**tan4 hui3**	na4 pu4	**hen4 shi4**	kuo4 de2	li4 su1

block11	block12	block13	block14	block15	block16	block17	block18	block19	block20
/t^h/	/ɲ/	/h/	/k'/	/ɬ/	/t^han/	/niŋ/	/hən/	/k'ʊŋ/	/ɬəu/
di4 lu3	yao4 zhu1	**hai4 ji4**	**kuan4 dui4**	ruan4 ji2	ti4 bing1	**ning4 ji2**	**hen4 zong1**	**kong4 tai4**	lie4 lin2
dai4 cai4	**nan4 hui4**	tai4 piao1	**kong4 ji2**	ru4 lun4	**tan4 ju2**	nang4 shi2	hai4 li3	kai4 zhong4	long4 ke4
tao4 kuo4	yang4 bai4	**han4 kao4**	pang4 li4	**lan4 qi2**	tu4 cong4	nao4 li2	**hen4 fu1**	ke4 rong2	**lao4 di4**
dao4 hu1	**ning4 lang2**	**hu4 zhang3**	**ku4 yuan2**	ren4 chi4	tie4 yong3	nai4 zong1	hao4 ni4	**kong4 dai4**	lu4 wen1
deng4 fa4	yan4 ding4	tan4 zan1	pai4 hua2	rao4 di1	tian4 quan1	**ning4 li2**	hu4 di4	**kong4 li3**	**lao 4 ku1**
dian4 sai4	yan4 xiang3	tou4 jue2	**ke4 tu2**	**liang4 gao1**	teng4 jie2	neng4 hao1	**hen4 ling2**	**kong4 tan2**	liu4 xi2

Table 3.9: Mandarin stimuli List B and D

3.2.2.3 Control of statistical features of English and Mandarin pseudo-word stimuli

For English and Mandarin stimuli in Experiment 1 and 2, I controlled the following 6 statistical features: (1) phonotactic probability of the initial phoneme, (2) phonotactic probability of the initial syllable, (3) phonotactic probability of the sum of phonemes, (4) phonotactic probability of the sum of biphones, (5) information-theoretic measures, and (6) neighborhood density.

3.2.2.3.1 Phonotactic probability Phonotactic probability refers to the frequency with which legal phonological segments or sequences of segments appear in a particular language (Jusczyk et al., 1994; Vitevitch & Luce, 2004). Previous studies have demonstrated that phonotactic probability serves as an important cue in speech recognition. Infants, for instance, can use phonotactic probability to discriminate between native sounds (Jusczyk et al., 1993; Jusczyk et al., 1994). For adults, Yip's research (2017) found that target words embedded in nonsense sound sequences with high transitional probability phoneme combinations were easier for listeners to identify. Additionally, studies comparing the processing of words and pseudo-words with varying phonotactic probabilities (Pitt & Samuel, 1995; Vitevitch & Luce, 1999) showed that listeners responded to pseudo-words with high phonotactic probability more quickly than to those with low phonotactic probability, indicating a facilitative effect of phonotactic probability for pseudo-words.

This suggests that variations in phonotactic probability can influence RTs in speech processing. Given that all the stimuli in this experiment are pseudo-words, it is essential to control the phonotactic probability of the segments in the stimuli to ensure they do not differ significantly.

The statistical features (1) to (4) refer to the phonotactic probability of segments (initial phoneme) and sequences of segments (initial syllable, sum of biphones, sum of phonemes) in the stimuli. Statistical features (1) to (4) for English stimuli are provided by the Phonotactic Probability Calculator from the University of Kansas (https://calculator.ku.edu). Features (5) and (6) were derived from the Database of Mitten Advanced Learner Dictio-

nary (https://ota.bodleian.ox.ac.uk/repository/xmlui/handle/20.500.12024/2469).

For Mandarin pseudo-word stimuli, all these features are provided by the Mandarin Lexical Database (CLD) from the University of Tübingen (http://www.mandarinlexicaldatabase.com). For English stimuli, I used the pseudo-word *telson* /ˈtɛl.sn̩/ as an example to calculate the statistical features of the initial phoneme and the stimulus *dagtil* /ˈdægtɪl/ as an example for the initial syllable. For Mandarin pseudo-word stimuli (with numbers indicating the tones), I used the stimulus *di4 lu3* /ti lu/ to calculate the statistical features of the initial phoneme, and the stimulus *ti4 bing1* /t^hi piŋ/ for the initial syllable.

The phonotactic probability (PP) of the initial phoneme is calculated by dividing the frequency (F) of the phoneme appearing as the initial phoneme /t/ by the total number of words in the corpus.

(1) PP of initial phoneme /t/ in English and Mandarin stimuli:

$$\text{PP}_{\text{/t/}} = \frac{\text{F}_{/t/}}{\text{Corpus}}$$

The PP of the initial syllable /dæg/ in English pseudo-word stimuli and /t^hi/ in Mandarin pseudo-word stimuli is calculated by dividing the frequency of each syllable appearing as the initial syllable by the total number of words in the corpus.

(2a) PP of initial syllable /dæg/ in English pseudo-word stimuli:

$$\text{PP}_{\text{/dæg/}} = \frac{\text{F}_{/dæg/}}{\text{Corpus}}$$

(2b)PP of initial syllable /t^hi/ in Mandarin pseudo-word stimuli:

$$\text{PP}_{\text{/t}^h\text{i/}} = \frac{\text{F}_{/t^h i/}}{\text{Corpus}}$$

The PP of the sum of phonemes is calculated by summing up the phonotactic probability of each phoneme.

(3a) PP of sum of phonemes /ˈtɛl.sn̩/ and /ˈdægtɪ l/ in English pseudo-word stimuli:

$$\text{Sum of Phonemes}_{\text{/tɛl.sn̩/}} = \frac{\text{F}_{/t/} + F_{/\text{ɛ}/} + F_{/l/} + F_{/s/} + F_{n}}{\text{Corpus}}$$

$$\text{Sum of Phonemes}_{\text{/dægtɪl/}} = \frac{\text{F}_{/d/} + F_{/\text{æ}/} + F_{/g/} + F_{/t/} + F_{/I/} + F_{/l/}}{\text{Corpus}}$$

(3b) PP of sum of phonemes /ti lu/ and /tʰi piŋ/ in Mandarin pseudo-word stimuli:

$$\text{Sum of Phonemes}_{\text{/ti lu/}} = \frac{\text{F}_{/t/} + F_{/i/} + F_{/l/} + F_{/u/}}{\text{Corpus}}$$

$$\text{Sum of Phonemes}_{\text{/tʰi piŋ/}} = \frac{\text{F}_{/t^{\text{h}}/} + F_{/i/} + F_{/p/} + F_{/i/} + F_{/\text{ŋ}/}}{\text{Corpus}}$$

The PP of sum of biphones is calculated by summing up the transitional phonotactic probability of each two neighboring phonemes.

(4a) PP of sum of biphones /ˈtɛl.sn̩/ and /ˈdægtɪl/ in English pseudo-word stimuli:

$$\text{Sum of Biphones}_{\text{/tɛl.sn̩/}} = \frac{\text{F}_{/t\text{ɛ}/} + F_{/\text{ɛ}l/+}\text{F}_{/ls/} + \text{F}_{/sn/}}{\text{Corpus}}$$

$$\text{Sum of Biphones}_{\text{/dægtɪl/}} = \frac{\text{F}_{/d\text{æ}/} + \text{F}_{/\text{æ}g/} + \text{F}_{/gt/} + \text{F}_{/tI/} + \text{F}_{/Il/}}{\text{Corpus}}$$

(4b) PP of sum of biphones /ti lu/ and /tʰi piŋ/ in Mandarin pseudo-word stimuli:

$$\text{Sum of Biphones}_{\text{/ti lu/}} = \frac{\text{F}_{/ti/} + \text{F}_{/il/} + \text{F}_{/lu/}}{\text{Corpus}}$$

$$\text{Sum of Biphones}_{\text{/tʰi piŋ/}} = \frac{\text{F}_{/t^{\text{h}}i/} + \text{F}_{/ip/} + \text{F}_{/pi/} + \text{F}_{/i\text{ŋ}/}}{\text{Corpus}}$$

3.2.2.3.2 Information-theoretic measures Statistical feature (5) encompasses information-theoretic measures, including Backward entropy, Backward conditional probability, and the Surprisal of an event. Information theory (Shannon, 1948) provides a framework for understanding uncertainty and information in language processing.

Backward entropy

The first type of information-theoretic measure considered is Shannon entropy of context, also known as backward entropy (BH). In this experiment, BH is calculated across the possible initial phonemes and syllables. BH describes the uncertainty associated with the first phoneme and syllable given the following context. High entropy indicates that many possible initial phonemes or syllables can precede the context, reflecting a high level of uncertainty. Conversely, low entropy suggests that only a few initial phonemes or syllables are typically used, indicating low uncertainty. For pseudo-words, entropy can be high because predictions about the segments are not concentrated around a single outcome. Pseudo-words with high entropy may lead to longer reaction times (Luce & Large, 2001).

To calculate the backward entropy (BH), we first need to obtain the joint probability (JP) of all possible initial phonemes or syllables along with their context. Two elements—"event" and "context"—need to be defined in BH. In English, the event consists of the initial phoneme and syllable, while the context refers to the segment adjacent to the initial phoneme and syllable, including the stress on the initial syllable. Since stress is always placed on the initial syllable in all stimuli, it is considered part of the context. For example, in English pseudo-word stimuli, the JP of the initial phoneme /t/ and its context /ˈχɛ/ is P(/ˈtɛ/), which indicates the probability of the initial phoneme /t/ appearing before /ɛ/ in a stressed syllable (χ indicates any possible initial phonemes). For the syllable, the JP of the initial syllable with stress /ˈdæg/ and its context /t/ is P(/ˈdægt/), which indicates the probability of /ˈdæg/ appearing before /t/ as an initial stressed syllable.

Therefore, for English pseudo-word stimuli, the calculation of the JP of the "event" and

"context" is:

(5a) The information-theoretic measures: Shannon entropy of context (BH) in English pseudo-word stimuli.

The joint probability of the initial phoneme /t/ ("event") is calculated first:

$$\text{JP}(/\text{'t}\varepsilon/) = \frac{\text{F}(/\text{'t}\varepsilon/)}{\text{Corpus}}$$

From the joint probability of the initial phoneme /t/, we can calculate BH of context /'$\chi\varepsilon$/. The χ refers to a possible initial phoneme that occurs preceding the given context /'$\chi\varepsilon$/. Since entropy concerns the uncertainty presented in a distribution of events, mathematically, entropy is the sum over a distribution of events (Siegelman et al., 2020). The BH for the initial phoneme is:

$$BH(/'\chi\varepsilon/) = -\sum[JP(/'\chi\varepsilon/) \times \log_2 JP(/'\chi\varepsilon/)]$$

Using the same calculation, the BH for the initial syllable is:

$$JP(/'d\text{æ}g.t/) = \frac{F(/'d\text{æ}g.t/)}{Corpus}$$

$$BH(/'\chi.t/) = -\sum[JP(/'\chi.t/) \times \log_2 JP(/'\chi.t/)]$$

In Mandarin pseudo-word stimuli, the *event* refers to the initial phoneme and syllable, with the initial syllable represented by the first character (C1). The *context* here refers to the segment adjacent to the initial phoneme and C1, considering tone 4. Since all C1s in the Mandarin pseudo-word stimuli are in tone 4, this tone is treated as part of the context. For the initial phoneme, the *context* is the second phoneme. For the initial syllable, the *context* is the first phoneme of the second syllable, which corresponds to the initial phoneme in C2. For example, in Mandarin pseudo-word stimuli, the JP of

the initial phoneme /t/ and its context /χi4/ is P(/χi4/), indicating the probability of the initial phoneme /t/ appearing before /i/ in a stressed syllable. For the syllable, the JP of the example C1 in tone 4 /t^hi4/ and its context /p/ is P(/t^hi4.p/), which indicates the probability of /t^hi4/ appearing before /p/ as C1 in tone 4.

Therefore, for Mandarin pseudo-word stimuli, the calculation of the JP of the *event* and *context* follows the same steps. The BH for the initial phoneme is:

$$JP(/ti4/) = \frac{F(/ti4/)}{Corpus}$$

$$BH(/\chi i4/) = -\sum[JP(/\chi i4/) \times \log_2 JP(/'\chi i4/)]$$

The BH for the initial syllable is:

$$JP(/t^h i4.p/) = \frac{F(/t^h i4.p/)}{Corpus}$$

$$BH(/\chi 4.p/) = -\sum[JP(/\chi 4.p/) \times \log_2 JP(/\chi 4.p/)]$$

Backward conditional probability

The second type of information-theoretic measure is backward conditional probability (BCP), which describes the probability of an event χ given subsequent events (Sun et al., 2018). A high BCP indicates that the frequency of χ and the context *C* are similar when χ consistently precedes *C*. Conversely, a low BCP suggests that the frequency of *C* is significantly higher than the frequency of χ preceding *C*. Prior studies have demonstrated that listeners utilize conditional probability as a contextual cue to predict upcoming words. High conditional probability facilitates word recognition in auditory speech perception (Frank & Willems, 2017; Lorenz & Tizón-Couto, 2019). Therefore, the BCPs of the target stimuli should be similar in magnitude.

The backward conditional probability I measured in English pseudo-word stimuli describes the probability of the initial phoneme and syllable given their context, specifically

the phoneme following the stressed initial phoneme and syllable.

In English pseudo-word stimuli, the backward conditional probability of the initial phoneme /t/ in a stressed syllable is calculated using the joint frequency of /'tɛ/, denoted as F(/'tɛ/), divided by the frequency of context F(/'χɛ/). The joint frequency of the initial phoneme F(/'tɛ/) represents the number of times the initial phoneme /t/ appears before /ɛ/ in a stressed syllable. In contrast, the frequency of context F(/'χɛ/) refers to the number of times the second phoneme /ɛ/ occurs in a stressed syllable.

In Mandarin pseudo-word stimuli, the backward conditional probability describes the probability of the initial phoneme and C1 given their context, which is the phoneme following the initial phoneme and C1 in tone 4. The backward conditional probability of the initial phoneme /t/ in tone 4 is calculated using the joint frequency of /ti4/, denoted as F(/ti4/), divided by the frequency of context of /χi4/ which is F(/χi4/). The joint frequency of the initial phoneme F(/ti4/) is the number of times that the initial phoneme /t/ appears before /i/ in tone 4. In contrast, the frequency of the context F(/χi4/) refers to the number of times the second phoneme /i/ occurs in tone 4.

(5b) The information-theoretic measures: backward conditional probability (BCP) in English pseudo-word stimuli. BCP of the initial phoneme is:

$$BCP(t|'\chi\varepsilon) = \frac{F(/'t\varepsilon/)}{F(/'\chi\varepsilon/)}$$

Using the same calculation, BCP of the initial syllable is:

$$BCP(d\text{æ}g|'\chi.t) = \frac{F(/'d\text{æ}g.t)}{F(/'\chi.t/)}$$

Using the same calculation, BCP of the initial phoneme in Mandarin pseudo-word stimuli is:

$$BCP(t|\chi i4) = \frac{F(/ti4/)}{F(/\chi i4/)}$$

BCP of the initial syllable in Mandarin pseudo-word stiimuli is:

$$BCP(t^{\mathrm{h}}i|\chi 4.p) = \frac{F(/t^{\mathrm{h}}i4.p/)}{F(/\chi 4.p/)}$$

Surprisal

The third type of information-theoretic measure is surprisal (S), calculated using backward conditional probability (BCP). Surprisal can be high because pseudo-words often involve unexpected sounds (Gwilliams & Davis, 2020). Surprisal is negatively correlated with BCP: a high BCP, indicating a high frequency of occurrence, results in low surprisal, and vice versa. Consequently, the surprisal of targets in the stimuli should also be of similar magnitude. In this experiment, surprisal quantifies the amount of information carried by the initial phoneme or syllable given a certain context. A high surprisal indicates a less predictable sequence, while a low surprisal indicates a more predictable sequence.

In English pseudo-word stimuli, S describes the initial /t/ appearing before /ɛ/ and the initial syllable /dæg/ appearing before /t/ in a stressed syllable. In Mandarin pseudo-word stimuli, S describes the initial /t/ appearing before /i/ and the initial syllable /ti/ appearing before /p/ in tone 4.

(5c) Information-theoretic measures: Surprisal (S).

In English pseudo-word stimuli:

$$S(/'t\varepsilon/) = -\log_2 BCP(t|'\chi\varepsilon)$$

$$S(/'d\text{æ}g.t/) = -\log_2 BCP(d\text{æ}g|'\chi.t)$$

In Mandarin pseudo-word stimuli:

$$S(/ti4/) = -\log_2 BCP(t|\chi i4)$$

$$S(/t^{\mathrm{h}}i4.p/) = -\log_2 BCP(t^{\mathrm{h}}i|\chi4.p)$$

3.2.2.3.3 Phonological neighborhood density The statistical feature (6) is phonological neighborhood density (PND). For Mandarin stimuli, this feature was derived from the Mandarin Lexical Database (CLD) (http://www.mandarinlexicaldatabase.com). For English stimuli, it was derived from the Database of the Mitten Advanced Learner Dictionary (https://ota.bodleian.ox.ac.uk/repository/xmlui/handle/20.500.12024/2469).

Prior studies have demonstrated that high neighborhood density has inhibitory effects on real-word speech processing, as competition among lexical representations slows down processing (Luce & Large, 2001; Pitt & Samuel, 1995; Vitevitch & Luce, 1999). Therefore, there should not be a significant difference in the number of neighbors for the target stimuli.

In non-tonal languages like English, a neighbor of a target word is defined as a word that differs from the target by the addition, deletion, or substitution of one phoneme in any position.

In Mandarin, a neighbor of a target word is defined as a word that differs from the target by one phoneme or one tone (Sun et al., 2018). The similarity neighborhood of a target word always includes the target word itself (Landauer & Streeter, 1973). The phonetic neighborhood density controlled here measures the number of neighbors of the target stimuli, including the target stimuli in the corpus.

3.2.2.3.4 Statistic features T-test for English and Mandarin stimuli The means of foils and targets for the six statistical features were compared within each list for stimuli in Experiments 1 and 2 using a paired t-test, revealing no significant differences. The t-test results for the English and Mandarin pseudo-word stimuli are presented in Table 3.10, indicating that the statistical features of foils and targets are comparable.

English list a and c					
Statistic Feature	t	df	*p*	MD	95%CI
PPp	-1.08	19	0.29	-0.01	-0.04, 0.01
PPs	-0.50	19	0.62	-1.3e-0.3	-0.01, 4.3e-03
SumPPp	1.66	19	0.11	0.03	-0.01, 0.06
SumPPb	0.10	19	0.92	2.5e-04	-0.01, 0.01
Entropy	0.91	19	0.38	0.21	-0.28, 0.71
BCP	-0.11	16	0.91	-5.48	-1.08, 9.72e-05
Surprisal	0.63	7	0.55	0.07	-0.19, 0.32
PND	0.46	19	0.65	0.04	-0.130.20

English list b and d					
Statistic Feature	t	df	*p*	MD	95%CI
PPp	-0.73	19	0.48	-0.01	-0.02, 0.01
PPs	-1.11	19	0.28	-0.01	-0.04, 0.01
SumPPp	1.26	19	0.22	0.02	-0.01, 0.06
SumPPb	0.03	19	0.97	9.1e-05	-0.01, 0.01
Entropy	0.74	19	0.47	0.17	-0.31, 0.66
BCP	0.35	16	0.73	3.1e-05	-1.6e-04, 2.2e-04
Surprisal	-0.23	7	0.82	-0.04	-0.50, 0.41
PND	0.35	19	0.73	0.02	-0.18, 0.16

Mandarin list a and c					
Statistic Feature	t	df	*p*	MD	95%CI
PPp	-1.80	19	0.09	-0.01	-0.03, 1.9e-03
PPs	0.17	19	0.87	115.79	-1326.28, 1557.86
SumPPp	0.74	19	0.47	38136.98	-69066.97, 145340.92
SumPPb	-0.55	19	0.59	-1.6e-04	-7.6e-04, 4.4e-04
Entropy	1.10	19	0.29	0.96	-0.87, 2.79
BCP	-0.08	19	0.94	-8.0e-04	-0.02, 0.02
Surprisal	0.71	19	0.49	0.49	-0.96, 1.94
PND	-1.12	19	0.28	-0.29	-0.820.25

Mandarin list b and d					
Statistic Feature	t	df	*p*	MD	95%CI
PPp	0.50	19	0.62	6463.51	-20747.23, 33674.25
PPs	0.03	19	0.98	17.49	-1418.24, 1453.21
SumPPp	0.91	19	0.37	45302.07	-58585.23, 149189.36
SumPPb	0.25	19	0.81	8.2e-05	-6.1e-04, 7.7e-04
Entropy	1.10	19	0.29	0.96	-0.87, 2.79
BCP	0.53	19	0.60	0.01	-0.02, 0.03
Surprisal	0.45	19	0.66	0.28	-1.03, 1.59
PND	-0.62	19	0.54	-0.17	-0.74, 0.40

Table 3.10: T-tests of statistical features between targets and foils for English and Mandarin pseudo-word stimuli, compared within-list. Rows indicate the information-theoretic values of interest, including the phonotactic probability of the first phoneme (PPp) and syllable (PPs), the sum of the phonotactic probability of phonemes (SumPPp) and biphones (SumPPb), entropy, backward conditional probability (BCP), surprisal, and phonological neighborhood density (PND). Columns show the t-statistics, degrees of freedom (df), p-value, mean difference between targets and foils (MD), and the 95% confidence interval (95%CI).

3.2.3 Participants

3.2.3.1 Experiment 1

The first group of participants consisted of 80 native speakers of English (referred to as English listeners). Forty of them completed the experiment offline in a lab on campus, while the other 40 completed it online. The second group included 80 native Mandarin speakers learning English as L2 (referred to as Mandarin listeners). Forty of them completed the experiment in a lab setting, and the other 40 completed it online. All participants were over the age of 18 and reported no hearing impairments. The experiment for each participant lasted approximately 12 minutes, and participants were financially compensated for their time (£5 for in-lab participation and £2 for online participation as they did not need to travel to the lab) [1].

Participants who completed the experiment in the lab were students at the University of Glasgow, recruited in the UK. Participants who completed the experiment online were recruited through mailing lists for undergraduate and postgraduate students at universities, as well as through social media and group chats for university students and English L2 learners. Mandarin listeners were required to have at least an intermediate level of proficiency in English (with IELTS scores above 5.5). Mandarin listeners who completed the experiment in the lab were all Chinese students studying abroad at the University of Glasgow and used English daily. Mandarin listeners who completed the experiment online were English L2 learners who were either studying abroad at the English-speaking Universities or had studied abroad but not necessarily in English-speaking environment at the time of experiment. They self-reported IELTS scores above 5.5.

3.2.3.2 Experiment 2

Experiment 2 recruited different participant groups compared to Experiment 1. The first group of participants consisted of 80 native speakers of Mandarin (referred to as Mandarin listeners), with 40 participants completing the experiment offline in a campus lab and another 40 completing it online. The second group comprised 56 native English speakers

[1] Online participants did not have access to information about the in-lab task or any differences in compensation, it is unlikely that the financial compensation affected task performance or motivation.

learning Mandarin as L2 (referred to as English listeners), with 28 participants completing the experiment offline in a campus lab and the other 28 completing it online. All participants were over the age of 18 and reported no hearing impairments. Participants were financially compensated £5 for in-lab participation and £2 for online participation.

Participants who completed the experiment in the lab were all students at the University of Glasgow, recruited in the UK. Participants who completed the experiment online were recruited through mailing lists for undergraduate and postgraduate students at universities, as well as through social media and group chats for university students and Mandarin L2 learners. English listeners were required to have at least an intermediate level of proficiency in Mandarin. Those who completed the experiment in the lab were all English students who had taken intermediate Mandarin lessons (Chinese 2) provided by the School of Modern Languages at University of Glasgow. English listeners who completed the experiment online were either former students who had taken the intermediate Mandarin lessons at the University of Glasgow or self-reported as Mandarin L2 learners with an intermediate level of proficiency, though not necessarily living in a Mandarin-speaking environment at the time of the experiment.

3.2.4 Procedure

All participants completed the experiment using an online link from the Gorilla online experimental platform (https://gorilla.sc/). In-lab participants also used the online link but in a physical lab setting. Each participant listened to 20 blocks, with each block containing 6 pseudo-words. Between each block, participants pressed the space bar to manually continue to the next block.

Experiment 1 Prior to the start of the experiment, listeners were informed via an audio recording in the instruction sections that two speakers, named "Gogo" and "Xixi." Gogo was a female Standard South British English speaker in her 40s, while Xixi was a female non-native English speaker with Chinese accent in her 20s.

Gogo and Xixi were set to have an immigrant background and were immigrants to a

fictional nation named "Nartten." Gogo produced the first ten blocks of stimuli, while Xixi produced the second ten blocks of stimuli. Both speakers used their original accents when speaking the language "Narttenian." Consequently, listeners were informed that English pseudo-words they heard were actually "Narttenian words," and they were not required to focus on these words' meanings.

For participants who completed the experiment without prior knowledge, they were only given general information about the two speakers, such as where they came from and their hobbies. They were told that there were two speakers: Gogo, who moved to Nartten from South UK, and Xixi, who moved to Nartten from a fictional nation called "Lardia Kingdom." The participants listened to a self-introduction from each speaker, given in their own voices and accents, which included some of their daily hobbies. Gogo's introduction, spoken in a British accent, was:

Hello, I'm Gogo. I'm the speaker of the first half of the experiment that you'll hear. In my spare time, I watch reality TV shows, although I probably should play football to get fit. My favorite food is fish and chips.

Xixi's introduction was:

Hello, I'm Xixi, I'm speaker of the second half experiment that you'll hear. In my spare time, I play volleyball and run marathon. I hate watching reality TV, Star Trek is great though. My favourite food is French onion soup.

In contrast, participants with prior knowledge were given different information. Instead of hobbies, they received information about the two speakers' spoken language and accent, including the native languages and accents of the speakers, as well as example words in an artificial accent for them to listen to. They were informed that Gogo produced Narttenian words with a standard British accent, while Xixi produced Narttenian words with a Lardian accent. Gogo's introduction was:

Hello, I'm Gogo, I moved to Nartten from South United Kingdom. I have to learn a

different language to live there. But my native language is British. I speak with standard British accent, I'm the speaker of block 1 to 10.

Xixi's introduction was:

Hello, I'm Xixi, I moved to Nartten from lardia Kingdom. I have to learn Narttenian language to live there. My native language is Lardian which is different from Narttenian. So I speak Narttenian with Lardian accent. I'm the speaker of block 11 to 20.

Participants with prior knowledge were also provided with text information about the features of the artificial accent they were going to listen to:

The following words are produced by Xixi with a Lardian accent, which is very heavy on the first sound.

Additionally, they were notified of the special pronunciation of the first sound the speakers produced and were asked to listen to words containing this special pronunciation: *language*, *native*, *block*, *kingdom* and *different*. These words contained the target word-initial phonemes in the artificial accent: voiceless alveolar lateral fricative /ɬ/, palatal /ɲ/, implosive /ɓ/, ejective /k'/, and retroflex /ɖ/.

After the prior knowledge section, each participant was given a randomized list of words to listen to. A word-initial phoneme or syllable target was displayed on the screen, and participants were asked to press the space bar on the keyboard to indicate "yes" if the word they heard began with the target showed on the screen. For example, with phoneme targets, participants would see the letter *t* appear on the screen. When they heard a word like telson, they would press the space bar. When they heard a word like dentawss, they did nothing. With syllable targets, participants would see the string of letters tar appear on the screen. When they heard a word like *tarlon*, they would press the space bar. When they heard a word like *larfo*, they did nothing. Before the trials began, participants were given 2 blocks of practice, with each block containing 4 practice trials, in total 8 practice trials. These practice trials were designed to ensure that participants clearly understood

the targets they needed to react to.

Experiment 2 The procedure of experiment 2 conducted in Mandarin was exactly identical as Experiment 1, and with instruction and prior knowledge given in Mandarin as well. Prior to the start of the experiment, two speakers were named "Weiwei (wēi wēi)" and "Xinxin (xīn xīn)." Weiwei was a female Standard native Mandarin speaker in her 40s, while Xixi was a female native Mandarin speaker with South accent in her 20s.

Weiwei and Xinxin were set to have an immigrant background and were immigrants to a fictional nation named "Nanyue (nán yuè)" . Weiwei produced the first ten blocks of stimuli, while Xinxin produced the second ten blocks of stimuli. Both speakers used their original accents when speaking the language "Nanyue language." Consequently, listeners were informed that Mandarin pseudo-words they heard were actually "Nanyue words."

For participants who completed the experiment without prior knowledge, they were told that: Weiwei, who moved to Nanyue from South China, and Xinxin, who moved to Nanyue from a fictional nation called "Liangzhu (liáng zhǔ)." Weiwei's introduction, spoken in a Standard Mandarin accent, was:

nín hǎo, wǒ jiào wēi wēi, zài běn cì shí yàn zhōng, dì yī dào dì shí mó kuài de cí huì shì wǒ shuō de. wǒ xǐ huān xì jù biǎo yǎn, yě ài kàn xiàn dài dū shì jù, wǒ zuì ài chī cháng fěn hé là wèi fàn, yīn cǐ wǒ xū yào tiào wǔ lái bǎo chí shēn cái.

Hello, I'm Weiwei. I'm the speaker of block 1 to 10. I enjoy acting in plays and watching modern TV dramas. I like rice noodle rolls and dried sausage rice, so I need to dance to keep fit.

Xinxin's introduction was:

nín hǎo, wǒ jiào xīn xīn, zài běn cì shí yàn zhōng, dì yī dào dì shí mó kuài de cí huì shì wǒ shuō de. wǒ suī rán bù tài ài kàn xì jù, dàn ài kàn zhōng guó gǔ dài gōng tíng jù, wǒ zuì ài bāo tāng chī mǐ fàn, yǒu shí huì qù hù wài bù xíng guān guāng.

Hello, I'm Xinxin. I'm the speaker of block 11 to 20. Although I don't really enjoy watching plays, I love watching ancient Chinese palace TV dramas. My favorite foods are soups and rice. Sometimes I go for outdoor walks and sightseeing.

Participants with prior knowledge were given the prior knowledge that Weiwei produced Nanyue words with standard Mandarin accent, while Xinxin produced Nanyue words with Liangzhu accent. Weiwei's introduction was:

nín hǎo, wǒ jiào wēi wēi, wǒ shì cóng zhōng guó nán bù lái dào nán yuè jū zhù de, yīn cǐ wǒ xū yào xué huì nán yuè yǔ cái néng zài nán yuè shēng huó, dàn shì wǒ de mǔ yǔ shì pǔ tōng huà, suǒ yǐ wǒ shuō nán yuè yǔ de shí hòu, yǒu pǔ tōng huà de kǒu yīn, dì yī dào dì shí mó kuài de cí huì shì wǒ shuō de.

Hello, I'm Weiwei, I moved to NanYue from South China. I have to learn a different language to live there. But my native language is Mandarin. I speak with standard Mandarin accent. I'm the speaker of block 1 to 10.

Xinxin's introduction was:

nín hǎo, wǒ jiào xīn xīn, wǒ shì cóng liáng zhǔ bān lái nán yuè jū zhù de, yīn cǐ wǒ xū yào xué huì nán yuè yǔ cái néng gèng hǎo de shēng huó, dàn shì wǒ de mǔ yǔ shì liáng zhǔ yǔ, suǒ yǐ wǒ shuō nán yuè yǔ de shí hòu, yǒu liáng zhǔ yǔ tè shū de kǒu yīn, dì shí yī dào dì èr shí bǎn kuài shì wǒ shuō de.

Hello, I'm Xinxin, I moved to Nanyue from Liangzhu. I have to learn Nanyue language to live there. My native language is Liangzhu which is different from Nanyue. So I speak Nanyue language with Liangzhu accent. I'm the speaker of block 11 to 20.

Participants with prior knowledge were also provided with text information about the features of the artificial accent they were going to listen to:

nín hǎo, wǒ shì xīn xīn, jiē xià lái de bǎn kuài shì wǒ shuō de, wǒ shuō nán yuè yǔ de shí

hòu dài yǒu liáng zhǔ yǔ de kǒu yīn.

Hi, I'm Xinxin, the following blocks you'll hear are from me. I still pronounce the Nanyue words with Liangzhu accent.

Information about the special pronunciation of the first sound the speakers produced was also given to participants, and they were asked to listen to words containing this special pronunciation: *tè shū "special"*, *huà yǔ 'speech'*, *nán yuè "Nanyue"*, *kǒu yīn "accent"* and *liáng zhǔ "liangzhu"*. These words contained the target word-initial phonemes in the artificial accent: aspirated retroflex plosive /ʈʰ/, palatal nasal /ɲ/, voiceless glottal fricative /h/, velar ejective /k'/, and voiceless alveolar lateral fricative /ɬ/.

Similar as Experiment 1, with phoneme targets, participants would see the letter such as *t* appear on the screen. When they heard a word such as *tào kuò*, they would press the space bar to indicate "yes." When they heard a word such as *dài cài*, they did nothing. With syllable targets, participants would see the string of letters *tan* appear on the screen. When they heard a word such as *tàn jú*, they would press the space bar, indicating "yes." When they heard a word such as *tì bìng*, they did nothing. Before the trials began, participants were also given 8 practice trials.

Chapter 4

Monitoring experiment results and discussion

4.1 Data pre-processing

4.1.1 Calculation of target offset

RTs for phoneme and syllable targets were measured and recorded. Listeners were presented with these targets and, upon hearing each target, they pressed a key on the keyboard to indicate detection. RTs were recorded by Gorilla from the beginning of each audio stimulus. Consequently, the actual RTs for listeners in response to phoneme and syllable targets were calculated by subtracting the target offset from the recorded reaction times.

This is because phoneme and syllable targets occur at different points in the stimulus. The initial phoneme of some targets has longer onsets, such as aspirated sounds, meaning their perceptual recognition point may be later than for other targets. Therefore, this adjustment ensures that the RTs accurately reflect the moment listeners were able to perceive the target sounds after it actually starts, rather than how much time has passed since the stimulus began. A complementary analysis using the original RTs measured from the beginning of the stimulus is provided in Appendix A.

The recorded RTs for the phoneme and syllable targets were adjusted by subtracting the respective target offsets for phonemes and syllables.

(1)

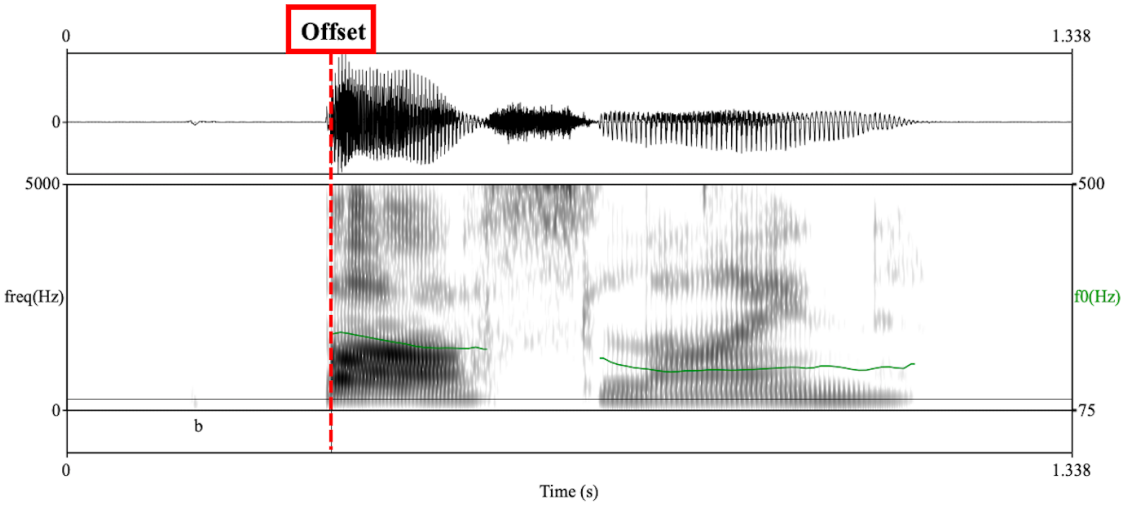
$$\mathrm{RT}_{phoneme} = \mathrm{RT}_{recorded} - \mathrm{Offset}_{phoneme}$$

(2)

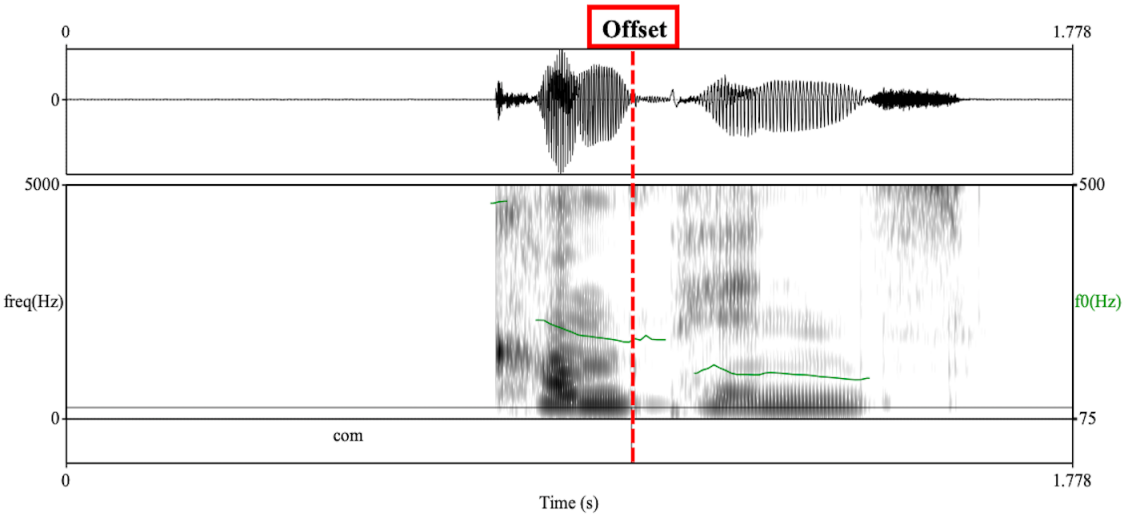
$$\mathrm{RT}_{syllable} = \mathrm{RT}_{recorded} - \mathrm{Offset}_{syllable}$$

The examples of phoneme and syllable target offset from both English and Mandarin stimuli are given in Figure 4.1 and 4.2.

The offsets of phoneme target /b/ in English pseudo-words *basline* and syllable target /kɔm/ in *compins* are showed in Figure 4.1. The offsets of phoneme target /t^h/ in Mandarin pseudo-words *tang4 yan2* and syllable target /hen/ in *hen4 teng2* are showed in Figure 4.2.

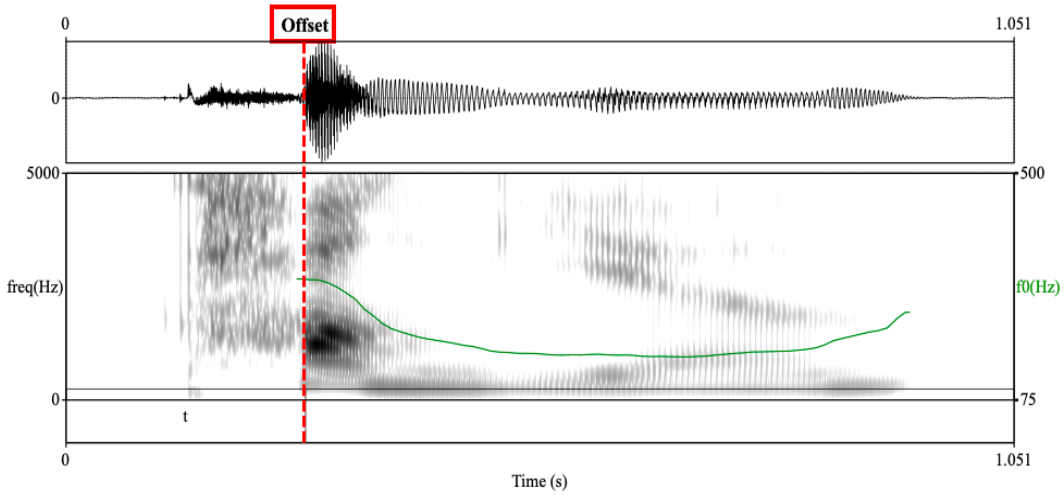


(a) Offset of phoneme /b/ in English pseudo-word *basline*

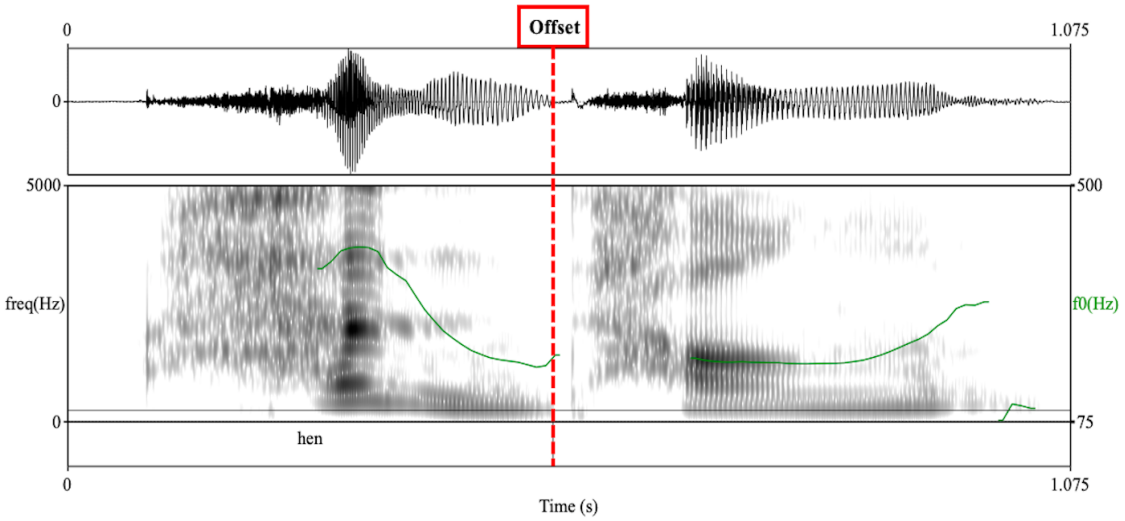


(b) Offset of syllable /kɔm/ in English pseudo-word *compins*

Figure 4.1: Phoneme and syllable target offsets of example words from English stimuli



(a) Offset of phoneme /t^h/ in Mandarin pseudo-word *tang4 yan2*



(b) Offset of syllable /hen/ in Mandarin pseudo-word *hen4 teng2*

Figure 4.2: Phoneme and syllable target offsets of example words from Mandarin stimuli

4.1.2 Calculation of masking effect

For an individual, the magnitude of masking effect (ME) was then calculated by subtracting this person's RTs of syllable targets from those of phoneme targets:

(3)

$$\mathrm{ME} = \mathrm{RT}_{phoneme} - \mathrm{RT}_{syllable}$$

There are three potential outcomes regarding the ME. If the ME is positive (> 0), it indicates that the RTs for phonemes are greater than the RTs for syllables, suggesting that listeners reacted faster to syllables than to phonemes. If the ME is close to 0, it suggests that the RTs for phonemes and syllables are similar. If the ME is negative (< 0), it indicates that the RTs for phonemes are smaller than the RTs for syllables, implying that listeners reacted faster to phonemes than to syllables.

Since the magnitude of masking effect is the difference between a phoneme and syllable, therefore, for each phoneme target matching with its corresponding syllable target (sharing the same word-initial phonemes), the label for the masking effect between these targets are showed in Tables 4.1 and 4.2. When calculating the masking effect, the result of the masking effect is grouped under each label for the subtraction between the phoneme targets and their corresponding syllable counterparts.

In the current monitoring Experiments 1 and 2, the masking effect was analyzed using a four-way repeated measures analysis of variance (ANOVA), aggregated both by participants (by-subject aggregation) and by stimuli (by-item aggregation). For by-subject aggregation, the masking effect for individual participant was calculated by subtracting the mean RTs of the syllable group from the mean RTs of the phoneme group across all conditions of critical trials. For by-item aggregation, the masking effect for each target label was calculated by subtracting the mean RTs of the syllable group from the mean RTs of the phoneme group across all participants in all conditions.

Label	Phoneme (Standard)	Phoneme (Artificial)	Syllable (Standard)	Syllable (Artificial)
d	/d/	/ɖ/	/den/	/ɖen/
n	/n/	/ɲ/	/ned/	/ɲed/
b	/b/	/ɓ/	/bɪs/	/ɓɪs/
k	/k/	/k'/	/kɔm/	/k'ɔm/
l	/l/	/ɬ/	/laɹ/	/ɬaɹ/

Table 4.1: English pseudo-word stimuli. Target labels for masking effect. *Standard* refers to the Standard Accent, *Artificial* refers the Artificial Accent.

Label	Phoneme (Standard)	Phoneme (Artificial)	Syllable (Standard)	Syllable (Artificial)
t	/t^h/	/$ʈ^h$/	/t^han/	/$ʈ^h$an/
l	/l/	/ɬ/	/lau/	/ɬəu/
n	/n/	/ɲ/	/niŋ/	/ɲiŋ/
h	/x/	/h/	/xən/	/hən/
k	/k^h/	/k'/	/k^hʊŋ/	/k'ʊŋ/

Table 4.2: Mandarin pseudo-word stimuli. Target labels for masking effect. *Standard* refers to the Standard Accent, *Artificial* refers the Artificial Accent.

4.2 Experiment 1: English pseudo-word stimuli

4.2.1 Accuracy of English and Mandarin listeners

The error rate is derived by dividing the number of critical trials that were not correctly responded (no button press was made) by the total number of critical trials. In Experiment 1, there are in total 80 participants for each group of listeners (English listeners and Mandarin listeners). Each participant responded to 50 critical trials. Thus, the total number of critical trials collected in data is 4000 each group of listeners .

Therefore, for English native listeners, there were 312 trials with incorrect responses and 3688 trials with correct responses. Their error rate was 7.8%. By contrast, for Mandarin listeners, there were 613 trials with incorrect responses and 3387 trials with correct responses. As a result, the error rate for Mandarin listeners was 15.3%.

These results indicate that, for English pseudo-word stimuli, English listeners have lower error rate compared to Mandarin listeners. The performance gap suggests that Mandarin listeners, being non-native listeners of English, faced greater challenges in recognizing the targets, likely due to unfamiliarity with English phonetics and phonology.

4.2.2 Results of Masking effect

4.2.2.1 Masking effect of English listeners

By-item analysis

When data was aggregated by item, the masking effect of English listeners is illustrated in Figure 4.3. The calculated statistics and interactions between variables in the ANOVA are showed in Table 4.3. The by-item analysis for English listeners showed neither significant main effect of Accent, Prior knowledge and Modality, nor significant interactions between them.

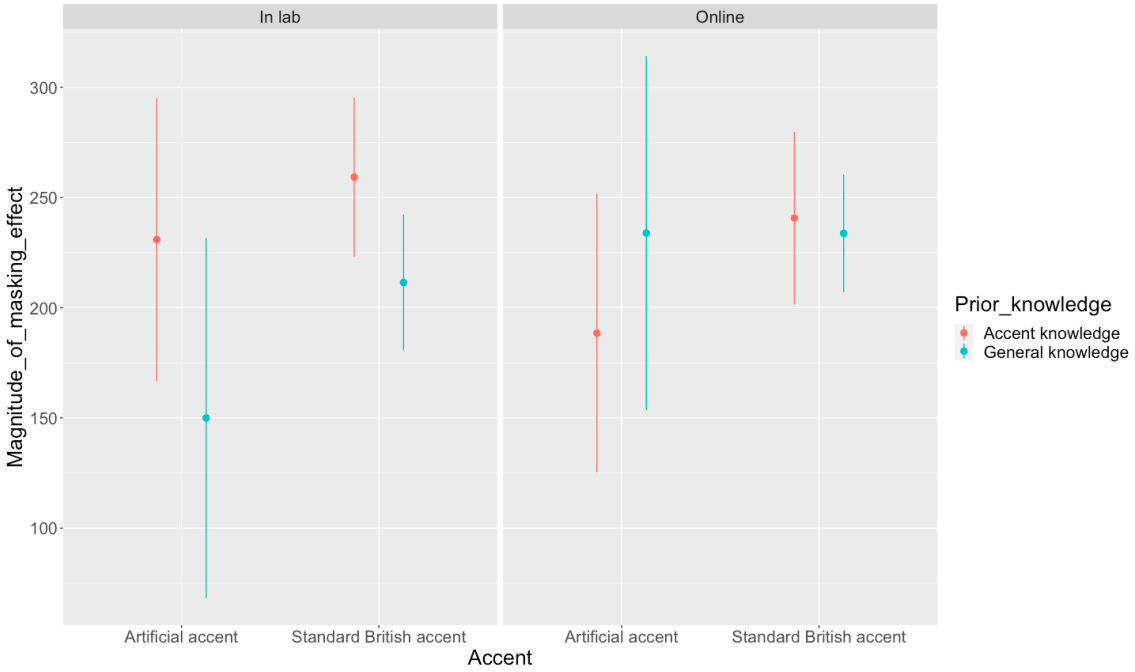


Figure 4.3: Experiment 1. English pseudo-word stimuli. Aggregated by-item. Masking effect for English listeners in milliseconds (ms) on y-axis. Conditions of *Accent* (Artificial accent/Standard British accent) are on x-axis. Fill color indicates the conditions of *Prior knowledge*(Accent knowledge/General knowledge). The facets indicate the conditions of *Modality* (In lab/Online).

Effect	DFn	DFd	F	p
pk	1	16	0.23	0.63
modality	1	16	0.06	0.81
accent	1	16	1.18	0.29
pk:modality	1	16	0.81	0.38
pk:accent	1	16	0.02	0.89
modality:accent	1	16	0.08	0.78
pk:modality:accent	1	16	0.43	0.52

Table 4.3: Experiment 1. English pseudo-word stimuli. Aggregated by-item. ANOVA of masking effect for English listeners. Prior knowledge is referred to as *pk*.

By-subject analysis

When data is aggregated by participants, masking effect of English listeners is illustrated in Figure 4.4, while the calculated statistics and interactions between variables in ANOVA are showed in Table 4.4. As showed by the Figure 4.4 and Table 4.4, there is no main effect for Modality ($F(1, 76) = 0.37$, $p = 0.55$). The main effect of Prior knowledge ($F(1, 76) = 3.94$, $p = 0.05$) is marginally significant. However, a significant interaction between prior knowledge and modality was observed ($F(1, 76) = 5.58$, $p < 0.05$).

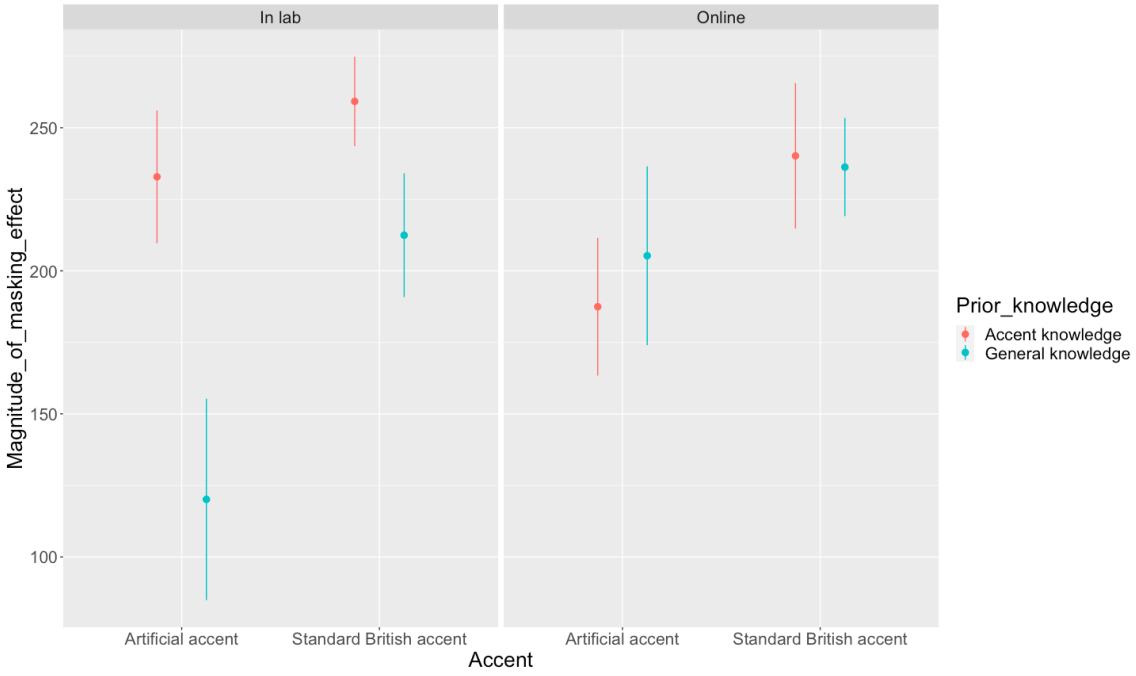


Figure 4.4: Experiment 1. English pseudo-word stimuli. Aggregated by-subject. Masking effect for English listeners in milliseconds (ms) on y-axis. Conditions of *Accent* (Artificial accent/Standard British accent) are on x-axis. Fill color indicates the conditions of *Prior knowledge*(Accent knowledge/General knowledge). The facets indicate the conditions of *Modality* (In lab/Online).

Effect	DFn	DFd	F	p
pk	1	76	3.94	0.05
modality	1	76	0.37	0.55
accent	1	76	8.95	< 0.01
pk:modality	1	76	5.58	< 0.05
pk:accent	1	76	0.43	0.52
modality:accent	1	76	0.27	0.61
pk:modality:accent	1	76	1.68	0.2

Table 4.4: Experiment 1. English pseudo-word stimuli. Aggregated by subject. ANOVA of masking effect for English listeners. Prior knowledge is referred to as *pk*.

Subsequent post-hoc t-test (Welch two independent samples t-test, *Prior knowledge* as between-subject variables) confirmed the interaction between Prior knowledge and Modality. When listeners completed the experiment in a lab setting, Prior knowledge had a significant effect ($t(66.6) = 3.09$, $p < 0.01$). Figure 4.4 shows that the application of Prior knowledge on accent (orange point ranges) increased the masking effect in a lab environment (left orange point ranges are higher than left blue point ranges). The changes in the masking effect are determined by the changes in the RTs to phoneme and syllable targets, as showed in Figure 4.5. In the lab setting, prior knowledge on accent decreased the RTs to syllable targets (left bottom blue point range), resulting in an increased distance between the RTs to phonemes and syllables, leading to an increased masking effect showed in the Figure 4.4. Figure 4.5 below displays the interaction between Prior knowledge and Modality, on the RTs to phoneme and syllable targets.

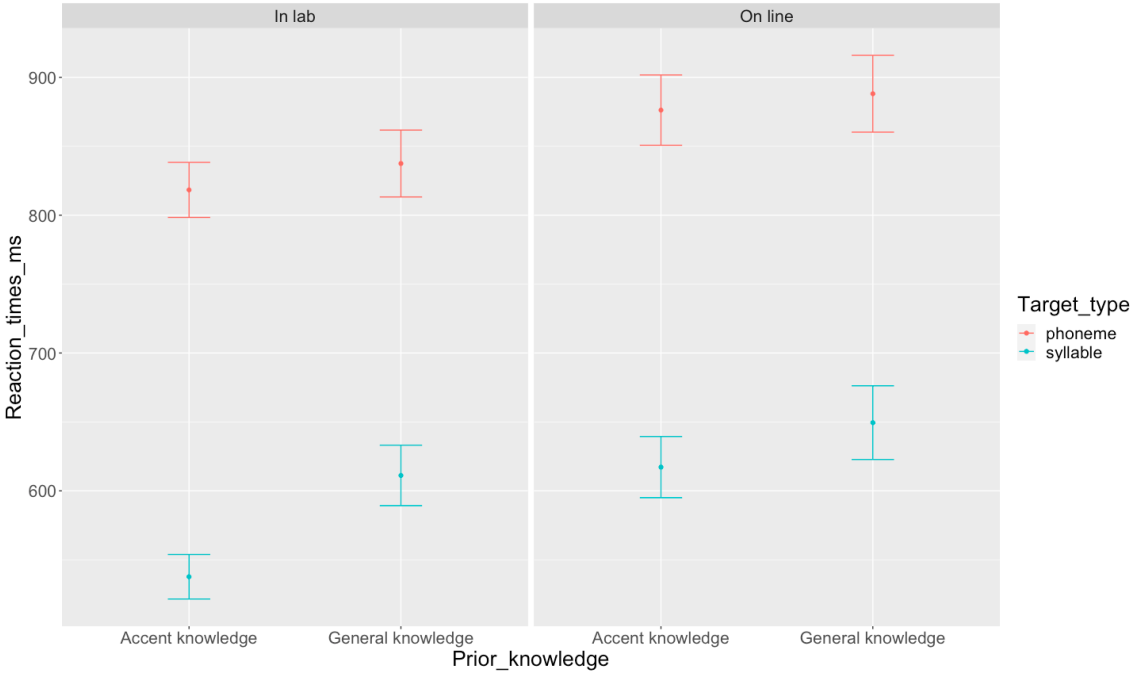


Figure 4.5: Experiment 1. English pseudo-word stimuli. RTs to phoneme and syllable targets for English listeners in milliseconds (ms) on y-axis, showing interaction between *Prior knowledge* and *Modality*.

There is a significant main effect of Accent as bottom-up information ($F(1, 76) = 8.95$, $p < 0.01$). As showed in Figure 4.4, there is a significant decrease in the masking effect when the artificial accent was applied. As showed in Figure 4.6 below, RTs to phoneme and syllable targets for English listeners reveal that the artificial accent slowed down the RTs

to both types of targets (left point ranges are higher than right point ranges), with RTs increasing more for syllables than for phonemes. This resulted in a smaller difference between the RTs to the two target types, leading to a smaller masking effect.

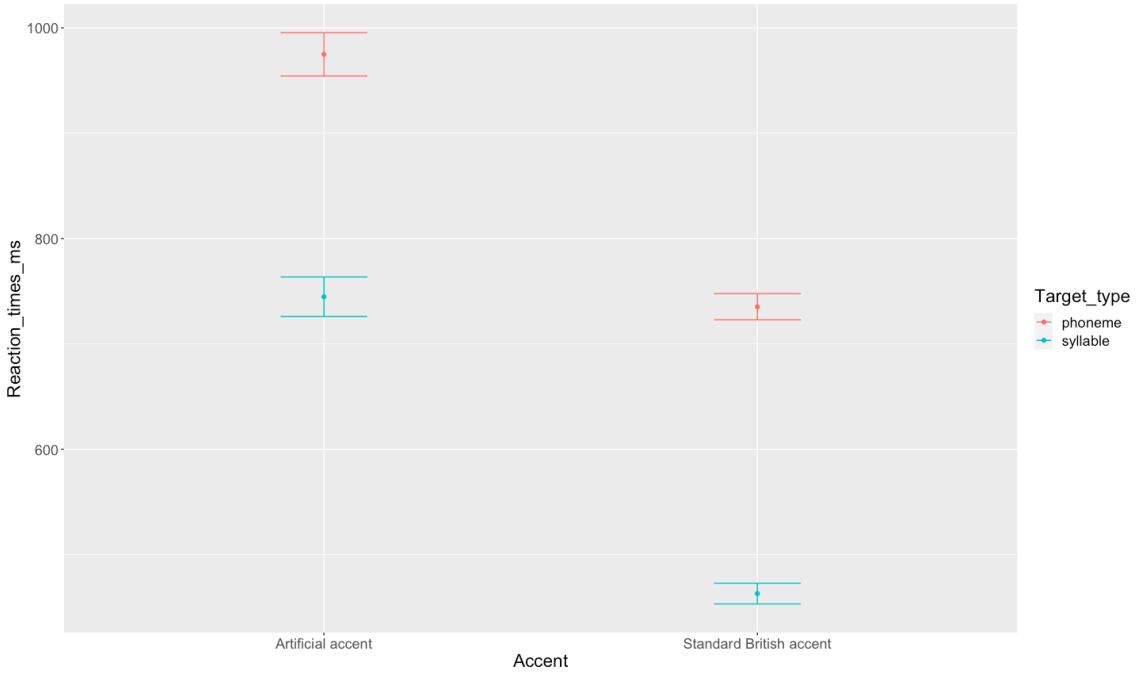


Figure 4.6: Experiment 1. English pseudo-word stimuli. RTs to phoneme and syllable targets for English listeners in milliseconds (ms) on y-axis, showing effect of *Accent*.

Further, there is no interaction between Accent and Prior knowledge ($F(1, 76) = 0.43$, $p = 0.52$), suggesting that the application of prior knowledge has little effect on English listeners' detection of targets with an artificial accent. There is no interaction between Modality and Accent ($F(1, 76) = 0.27$, $p = 0.61$), which means that the artificial accent had a significant effect on the masking effect regardless of whether listeners were in the lab or at their own locations. Additionally, the three-way interaction between Prior knowledge, Modality, and Accent was also insignificant for English listeners ($F(1, 76) = 1.68$, $p = 0.2$).

Figure 4.7 illustrates the RTs of the phonemes and syllables demonstrated by English native listeners. Overall, as showed by Figure 4.7, the results indicate that RTs to syllable targets were consistently faster than RTs to phoneme targets, consistent with previous research (Goldinger & Azuma, 2003; Savin & Bever, 1970) that larger units are responded faster than smaller units, demonstrating a typical masking effect. As the top-down infor-

mation, Prior knowledge did not have a main effect on the masking effect, but interacted with Modality, significantly increasing the magnitude of masking effect in the lab setting, by slowing down RTs to syllables more than phonemes. By contrast, Accent as bottom-up information did not interact with Modality but had a main effect on masking effect, with the artificial accent slowing down RTs to syllables more than phonemes, leading to a smaller difference between RTs to phonemes and syllables and creating the smaller masking effect.

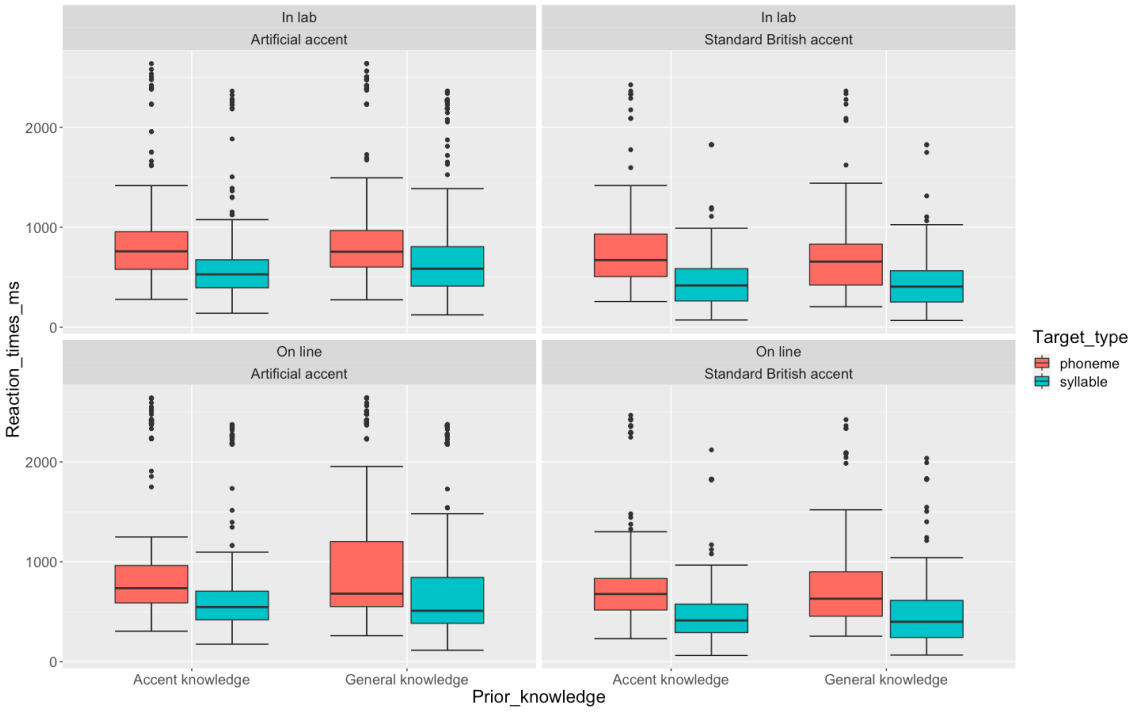


Figure 4.7: Experiment 1. English pseudo-word stimuli. RTs to phonemes and syllables for English listeners in milliseconds (ms) on the y-axis and the conditions of *Prior knowledge* (Accent knowledge/General knowledge) on the x-axis. Fill color indicates *Target type* (Phoneme/Syllable). Facets indicate the conditions of *Accent* (Artificial accent/Standard British accent) and conditions of *Modality* (In lab/Online).

4.2.2.2 Masking effect of Mandarin listeners

By-item analysis

When data was aggregated by item, the masking effect of Mandarin listeners is illustrated in Figure 4.8. The calculated statistics and interactions between variables in the ANOVA are showed in Table 4.5. As non-native listeners, similar to the results from English listeners, Mandarin listeners showed neither significant main effects for Prior knowledge, Accent and Modality nor interactions between them in the by-item analysis.

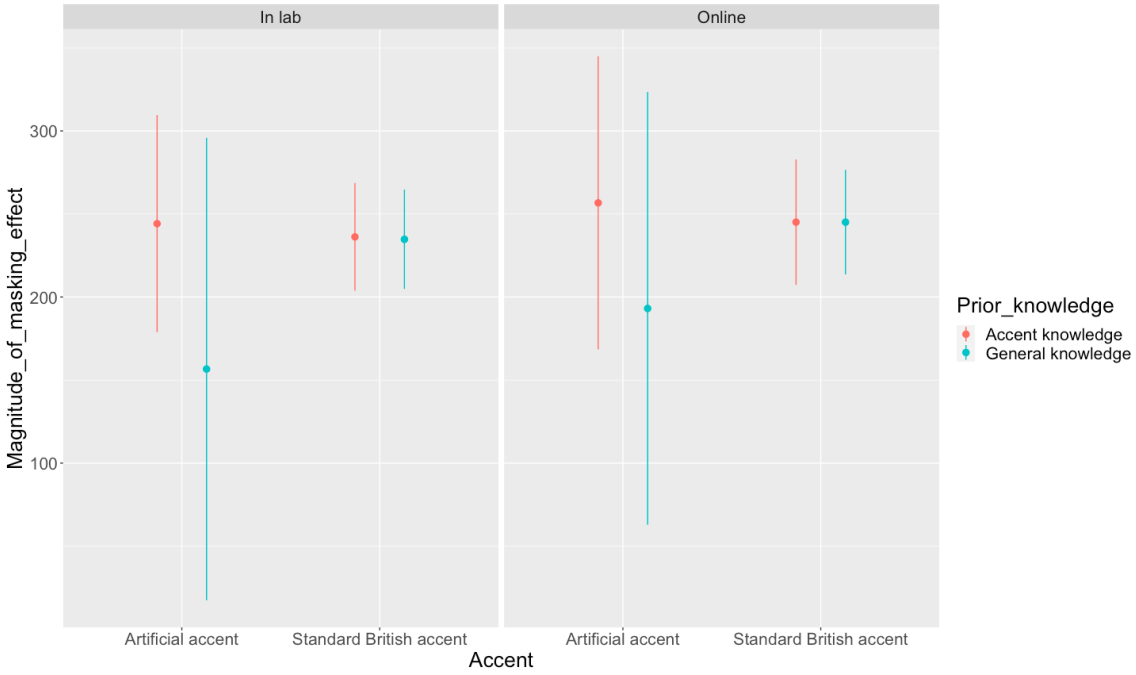


Figure 4.8: Experiment 1. English pseudo-word stimuli. Aggregated by-item. Masking effect of Mandarin listeners in milliseconds (ms) on y-axis. Conditions of *Accent* (Artificial accent/Standard British accent) are on x-axis. Fill color indicates the conditions of *Prior knowledge*(Accent knowledge/General knowledge). The facets indicate the conditions of *Modality* (In lab/Online).

Effect	DFn	DFd	F	p
pk	1	16	0.40	0.54
modality	1	16	0.08	0.78
accent	1	16	1.26	0.62
pk:modality	1	16	0.01	0.92
pk:accent	1	16	0.47	0.50
modality:accent	1	16	0.02	0.89
pk:modality:accent	1	16	0.01	0.92

Table 4.5: Experiment 1. English pseudo-word stimuli. Aggregated by-item. ANOVA of masking effect for Mandarin listeners. Prior knowledge is referred to as *pk*.

By-subject analysis

When data is aggregated by participants, masking effect of Mandarin listeners is illustrated in Figure 4.9, while the calculated statistics and interactions between variables in ANOVA are showed in Table 4.6. Modality for Mandarin neither has a main effect (F(1, 76) = 0.05, p = 0.82) nor interacts with Prior Knowledge (F(1, 76) = 0.22, p = 0.64) or Accent (F(1, 76) = 0.07, p = 0.80). This suggests that Mandarin listeners consistently experienced a masking effect regardless of the experiment location, whether in a lab or in their own environments.

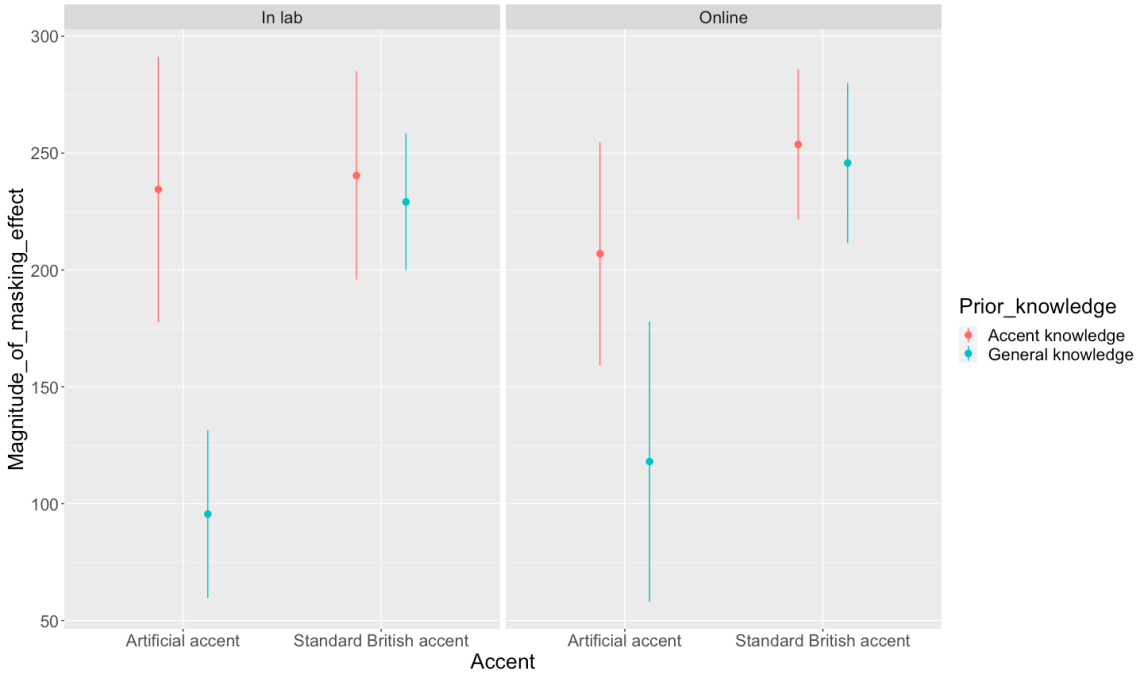


Figure 4.9: Experiment 1. English pseudo-word stimuli. Aggregated by-subject. Masking effect of Mandarin listeners in milliseconds (ms) on y-axis. Conditions of *Accent* (Artificial accent/Standard British accent) are on x-axis. Fill color indicates the conditions of *Prior knowledge*(Accent knowledge/General knowledge). The facets indicate the conditions of *Modality* (In lab/Online).

Effect	DFn	DFd	F	p
pk	1	76	4.75	$<$0.05
modality	1	76	0.05	0.82
accent	1	76	5.48	$<$0.05
pk:modality	1	76	0.22	0.64
pk:accent	1	76	2.42	0.12
modality:accent	1	76	0.07	0.80
pk:modality:accent	1	76	0.12	0.73

Table 4.6: Experiment 1. English pseudo-word stimuli. Aggregated by-subject. ANOVA of masking effect for Mandarin listeners. Prior knowledge is referred to as *pk*.

Mandarin listeners showed no three-way interaction between Accent, Prior knowledge and Modality, as showed in Table 4.6. However, the effect of Prior knowledge on the masking effect was weakly significant for Mandarin native listeners ($F(1, 76) = 4.75$, $p < 0.05$). In Figure 4.9, with prior knowledge on accent applied, the masking effect increased compared to when only general knowledge was applied (orange point ranges are higher than blue point ranges). This is because, as showed by the RTs to phonemes and syllables in Figure 4.9, prior knowledge on accent reduced the RTs to syllables more than to phonemes, resulting in an increased distance between RTs to phonemes and syllables, leading to an increased masking effect. Figure 4.10 displays the effect of Prior knowledge, showing how RTs to phonemes and syllables affect the changes in the masking effect showed in Figure 4.9.

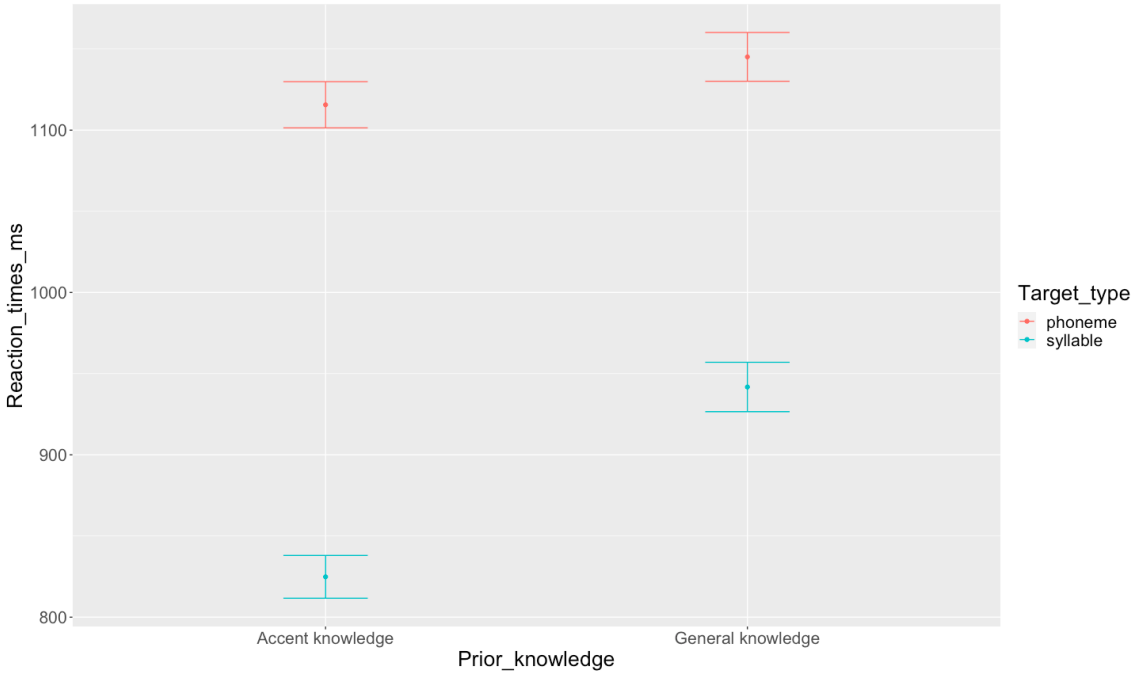


Figure 4.10: Experiment 1. English pseudo-word stimuli. RTs to phoneme and syllable targets for Mandarin listeners in milliseconds (ms) on y-axis, showing effect of *Prior knowledge*.

Moreover, there was a significant main effect of Accent on the masking effect for Mandarin listeners ($F(1, 76) = 5.48$, $p < 0.05$). The artificial accent, as bottom-up information, significantly reduced the masking effect. Compared to English listeners ($F(1, 76) = 8.95$, $p < 0.01$), the effect of Accent was less significant for Mandarin listeners, indicating that the artificial accent did not reduce the masking effect for Mandarin listeners as drasti-

cally as it did for English listeners. Figure 4.11 below shows how the changes in RTs to syllables and phonemes result in the decrease in the masking effect showed in Figure 4.9. Showed in Figure 4.11, the artificial accent slowed down RTs for Mandarin listeners to both types of targets, with a greater increase in RTs for syllables than for phonemes, resulting in a smaller masking effect compared to when the standard British accent was applied. This pattern is similar to that observed in English listeners.

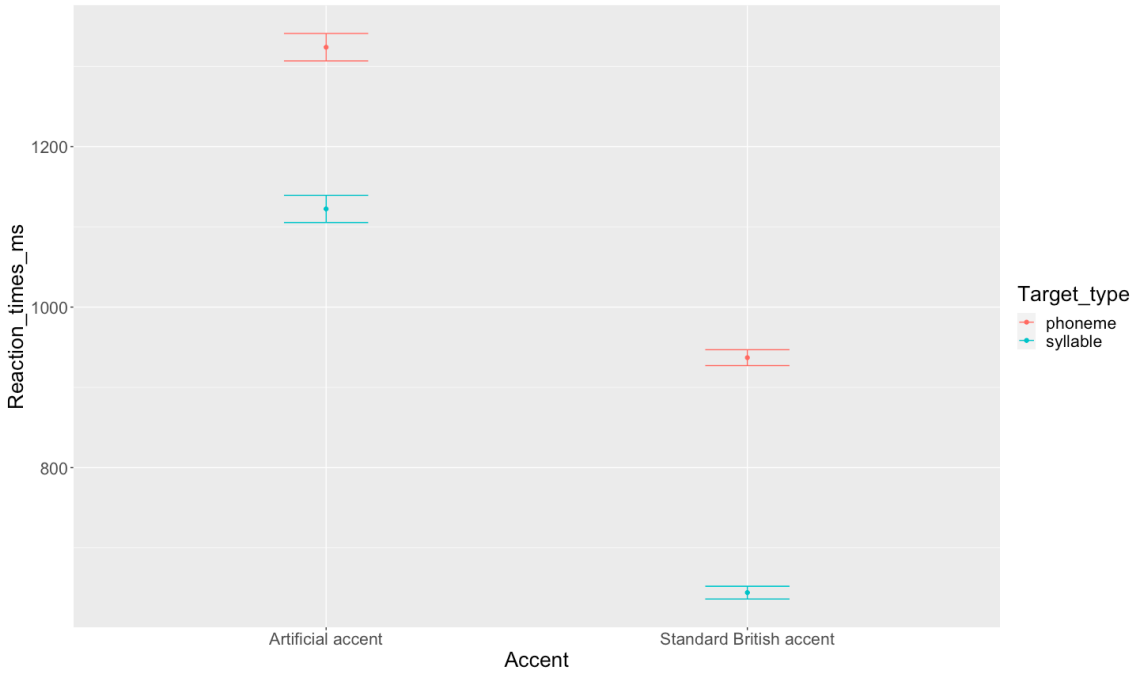


Figure 4.11: Experiment 1. English pseudo-word stimuli. RTs to phoneme and syllable targets for Mandarin listeners in milliseconds (ms) on y-axis, showing effect of *Accent*.

Figure 4.12 shows RTs for phonemes and syllables of Mandarin listeners. Similar to native English listeners, Mandarin listeners also displayed a typical masking effect, with RTs to syllable targets consistently faster than those to phoneme targets. As top-down information, prior knowledge has a weak main effect on the masking effect. RTs to syllables decreased when participants were given prior knowledge of the artificial accent, resulting in a greater masking effect than when only general knowledge was applied. Accent, as bottom-up information, also has a main effect on the masking effect. RTs to syllables increased more than RTs to phonemes under the artificial accent, resulting in a smaller masking effect than when only the standard accent was applied.

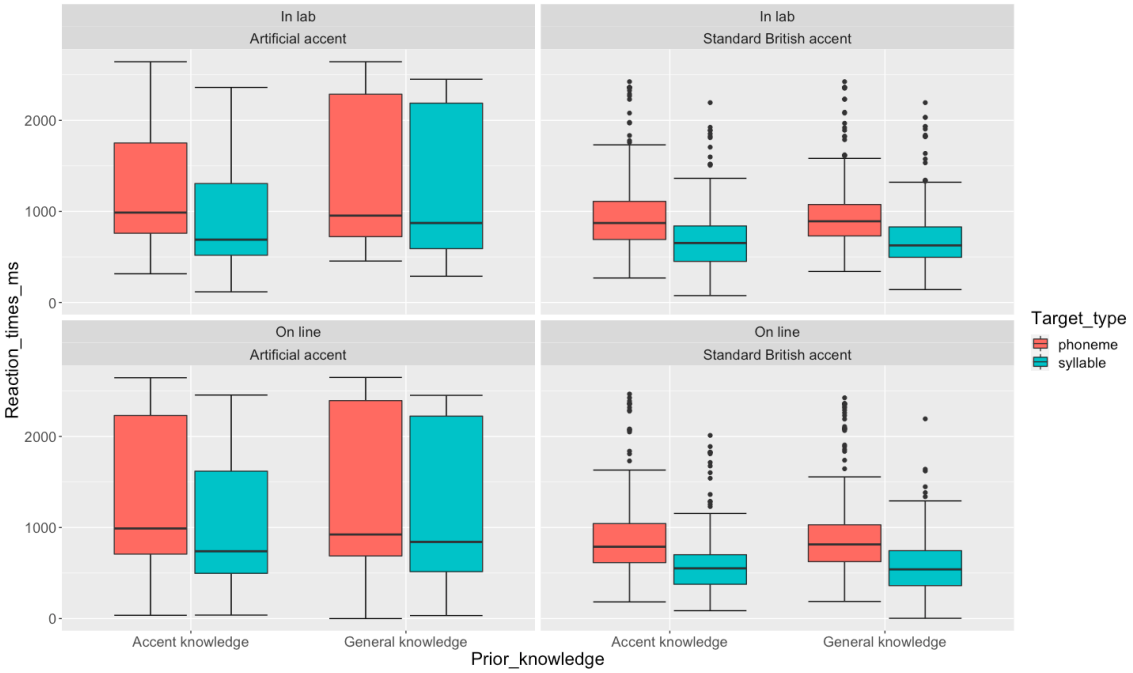


Figure 4.12: Experiment 1. English pseudo-word stimuli. RTs to phonemes and syllables for Mandarin listeners in milliseconds (ms) on the y-axis and the conditions of *Prior knowledge* (Accent knowledge/General knowledge) on the x-axis. Fill color indicates *Target type* (Phoneme/Syllable). Facets indicate the conditions of *Accent* (Artificial accent/Standard British accent) and conditions of *Modality* (In lab/Online).

4.2.3 Summary

In Experiment 1, using English pseudo-word stimuli, native Mandarin listeners showed a similar pattern to native English listeners. For both groups, a typical masking effect was observed, suggesting that even with manipulations of bottom-up and top-down information at the phoneme level, listeners consistently reacted faster to syllable targets than to phoneme targets.

A by-item analysis showed few significant effects, which were only observed in the by-subject analysis. This could suggest that the variability between items is relatively low compared to the variability between subjects. The stimuli follow a consistent and similar phonological structure and pattern. Individual differences among participants, such as L2 proficiency, cognitive load, and settings and surroundings affecting participants, etc., are driving the observed effects, rather than differences between the items.

A by-subject analysis showed that, regarding the impact of bottom-up information, both groups of listeners exhibited a reduced masking effect when an artificial accent was applied. This reduction was due to a greater increase in RTs to syllables than to phonemes, which is opposite to the predictions. This suggests that both groups of listeners faced challenges in identifying phoneme and syllable targets under the artificial accent, showing increased RTs to both, but syllables were more sensitive to the manipulation of bottom-up information than phonemes. Moreover, the effect of Accent was more significant among English listeners ($F(1, 76) = 8.95$, $p < 0.01$) than among Mandarin listeners ($t(469.34) = -2.10$, $p < 0.05$). This suggests that English listeners exhibited greater fluctuation in the masking effect than Mandarin listeners, indicating higher perceptual flexibility under the impact of bottom-up information compared to Mandarin listeners.

Regarding the impact of top-down information, both groups of listeners showed an increased masking effect, resulting from a greater decrease in RTs for syllables than for phonemes. This indicates that prior knowledge of the artificial accent mitigated the difficulty caused by the artificial accent, with syllables benefiting more from the top-down information. However, English listeners showed an interaction between Prior Knowledge

and Modality, with the effect of Prior Knowledge being significant in the lab setting. This could suggest that participants pay more attention to top-down information in the lab setting than in their own environments, as outlined in the preceding section, where it was noted that prior knowledge significantly increases the masking effect. Further, both groups of listeners showed no interaction between Accent and Prior Knowledge, indicating that bottom-up and top-down information have independent impacts.

Overall, under an English pseudo-word context, Mandarin listeners showed a generally comparable perceptual flexibility to English listeners under the effect of Prior knowledge and Accent. However, under the impact of bottom-up information, compared to Mandarin listeners ($t(469.34) = -2.10$, $p < 0.05$), English listeners ($F(1, 76) = 8.95$, $p < 0.01$) showed a more substantial fluctuation in the size of the masking effect, resulting from greater changes in RTs to syllables than to phonemes. This suggests that English listeners could flexibly give more focus on syllables than phonemes in target identification under the impact of artificial accent.

4.3 Experiment 2: Mandarin pseudo-word stimuli

4.3.1 Accuracy of Mandarin listeners and English listeners

Experiment 2 used Mandarin pseudo-word stimuli, in which there were 80 Mandarin listener participants. Each responded to 50 critical trials, resulting in a total of 4,000 trials. Mandarin native listeners had a total of 442 incorrect responses out of 4000 critical trials. Therefore, Mandarin native listeners had an error rate of 11.05%.

There were 56 English listener participants, each of whom responded to 50 critical trials, resulting in a total of 2,800 trials. English native listeners showed a comparable error rate to Mandarin listeners, with a total of 289 incorrect responses and 2,611 correct responses. Thus, English listeners had an error rate of 10.3%.

4.3.2 Results of masking effect

4.3.2.1 Masking effect of Mandarin listeners

By-item analysis

When data was aggregated by item, the masking effect of Mandarin listeners is illustrated in Figure 4.13. The calculated statistics and interactions between variables in the ANOVA are showed in Table 4.7. The by-item analysis for Mandarin listeners showed neither significant main effects for Accent, Prior knowledge and Modality nor significant interactions between.

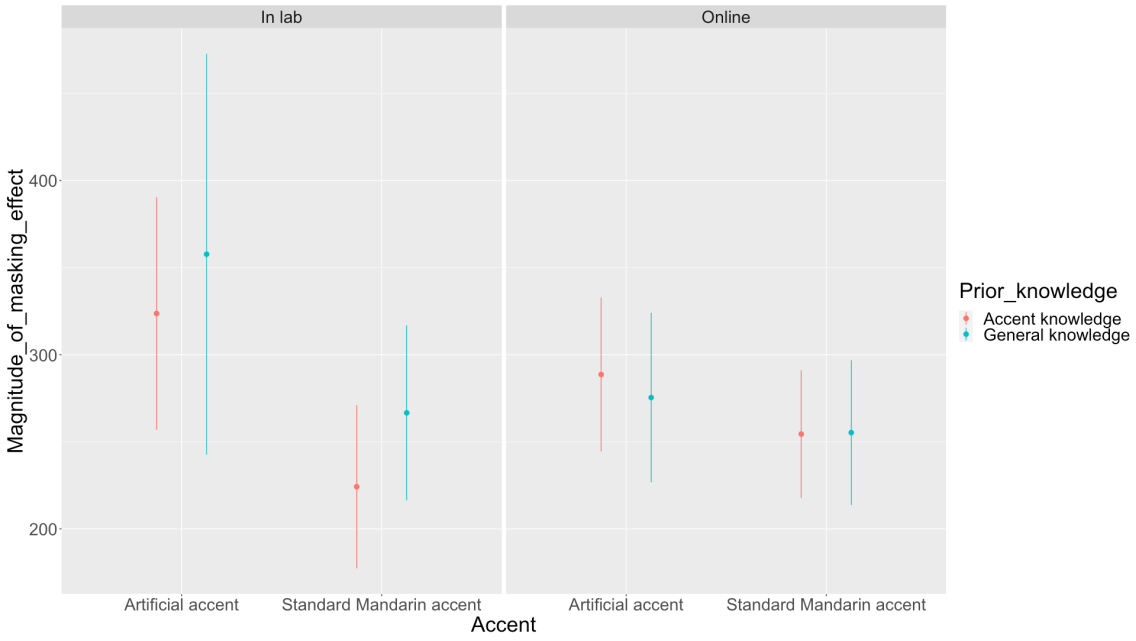


Figure 4.13: Experiment 2. Mandarin pseudo-word stimuli. Aggregated by-item. Masking effect of Mandarin listeners in milliseconds (ms) on y-axis. Conditions of *Accent* (Artificial accent/Standard Mandarin accent) are on x-axis. Fill color indicates the conditions of *Prior knowledge*(Accent knowledge/General knowledge). The facets indicate the conditions of *Modality* (In lab/Online).

Effect	DFn	DFd	F	p
pk	1	16	0.12	0.73
modality	1	16	0.27	0.61
accent	1	16	3.51	0.08
pk:modality	1	16	0.15	0.70
pk:accent	1	16	0.03	0.86
modality:accent	1	16	1.07	0.32
pk:modality:accent	1	16	0.002	0.96

Table 4.7: Experiment 2. Mandarin pseudo-word stimuli. Aggregated by-item. ANOVA of masking effect for Mandarin listeners. Prior knowledge is referred to as *pk*.

By-subject analysis

Figure 4.14 and Table 4.8 illustrate the masking effect of native Mandarin listeners, along with the statistical calculations and variable interactions conducted through ANOVA. Neither main effect of Accent, Prior knowledge and Modality, nor the interactions were showed.

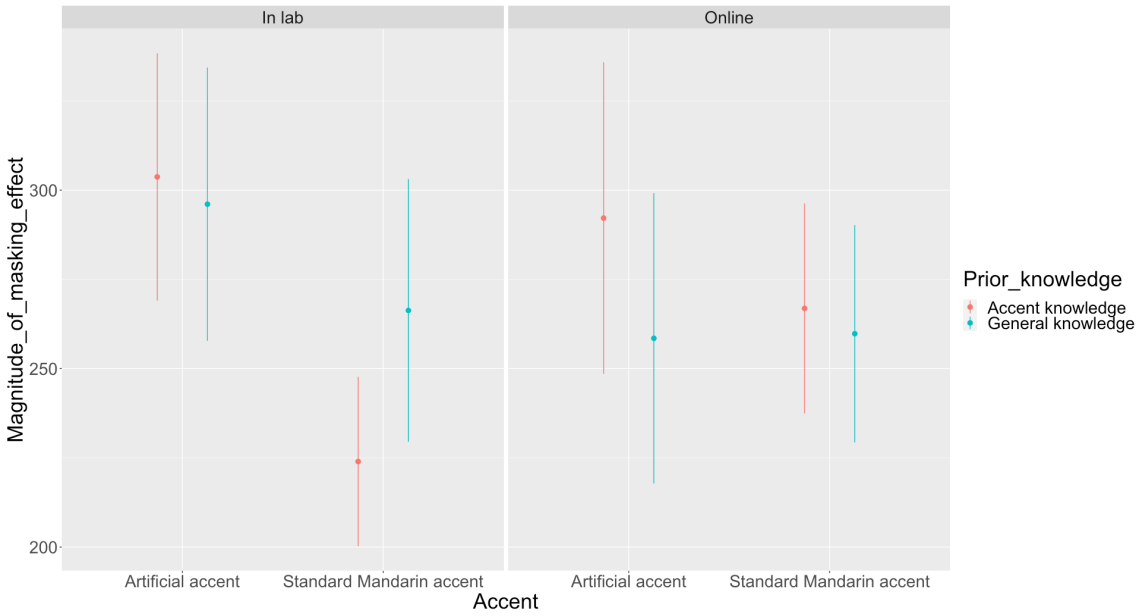


Figure 4.14: Experiment 2. Mandarin pseudo-word stimuli. Aggregated by-subject. Masking effect of Mandarin listeners in milliseconds (ms) on y-axis. Conditions of *Accent* (Artificial accent/Standard Mandarin accent) are on x-axis. Fill color indicates the conditions of *Prior knowledge*(Accent knowledge/General knowledge). The facets indicate the conditions of *Modality* (In lab/Online).

Effect	DFn	DFd	F	p
pk	1	76	0.01	0.92
modality	1	76	0.07	0.79
accent	1	76	2.73	0.10
pk:modality	1	76	0.28	0.60
pk:accent	1	76	0.97	0.33
modality:accent	1	76	0.85	0.36
pk:modality:accent	1	76	0.036	0.85

Table 4.8: Experiment 2. Mandarin pseudo-word stimuli. Aggregated by subject. ANOVA of masking effect for Mandarin listeners. Prior knowledge is referred to as *pk*.

Modality has no main effect (F(1, 76) = 0.07, p = 0.79) nor interaction with Accent (F(1, 76) = 0.85, p = 0.36) or Prior Knowledge (F(1, 76) = 0.28, p = 0.60). This suggests that the masking effect was not modulated by modality or its interactions. Moreover, Accent as bottom-up information (F(1, 76) = 2.73, p = 0.10) and Prior Knowledge as top-down information (F(1, 76) = 0.01, p = 0.92) do not have significant main effects on the masking effect for Mandarin listeners. Further, no significant interaction was found between Accent and Prior knowledge (F(1, 76) = 0.97, p = 0.33), indicating the independent effect of these two informational cues on the masking effect.

Figure 4.15 illustrates the RTs of native Mandarin listeners. As showed in Figure 4.15, using Mandarin pseudo-word stimuli, the results indicate that RTs to syllables were faster than RTs to phonemes, demonstrating a typical masking effect. Overall, the findings suggest that Prior Knowledge had neither a main effect nor an interaction with Modality. Additionally, Accent failed to interact with Modality and had little impact on the masking effect. Furthermore, no interaction was observed between Accent and Prior Knowledge, indicating that Mandarin listeners utilized these information sources independently.

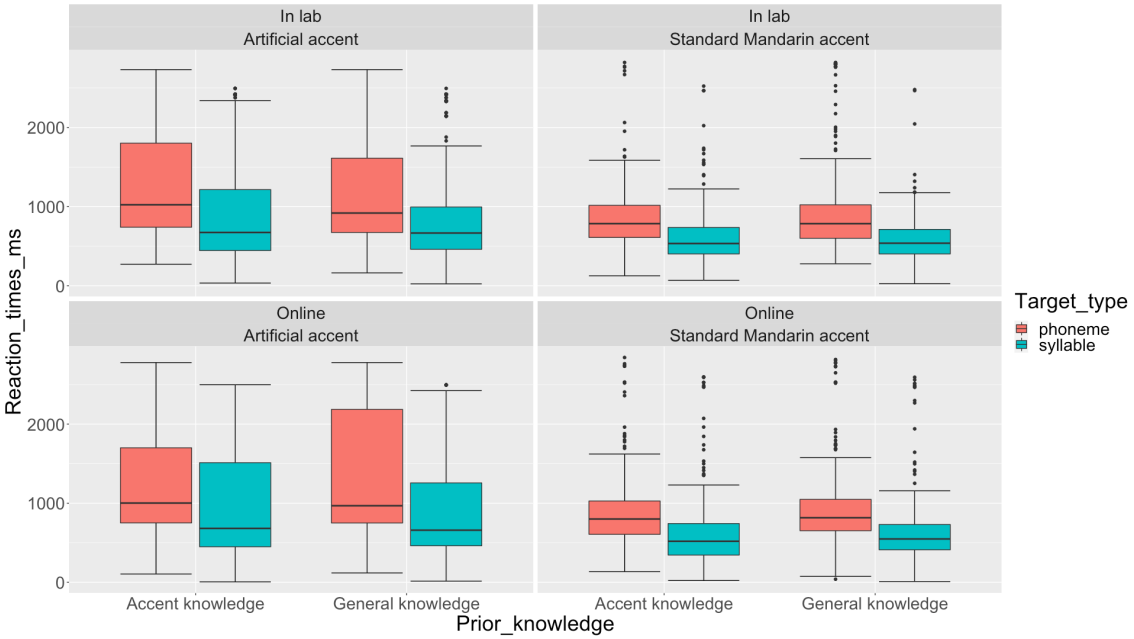


Figure 4.15: Experiment 2. Mandarin pseudo-word stimuli. RTs to phonemes and syllables for Mandarin listeners in milliseconds (ms) on the y-axis and the conditions of *Prior knowledge* (Accent knowledge/General knowledge) on the x-axis. Fill color indicates *Target type* (Phoneme/Syllable). Facets indicate the conditions of *Accent* (Artificial accent/Standard Mandarin accent) and conditions of *Modality* (In lab/Online).

4.3.2.2 Masking effect of English listeners

By-item analysis

When data was aggregated by item, the masking effect of English listeners is illustrated in Figure 4.16. The calculated statistics and interactions between variables in the ANOVA are showed in Table 4.9. Neither Accent nor Modality has main effect, but Accent has an interaction with Modality ($F(1, 16) = 5.56$, $p < 0.05$). However, the post-hoc t-test (paired t-test, *Accent* as within-item variable) showed that Accent has no significant effect neither when listeners completed experiments in a lab setting ($t(9) = 1.96$, $p = 0.08$) or online ($t(9) = -1.61$, $p = 0.14$). This further indicates that, similar as the previous by-item analysis, Accent has little effect when interacting with Modality.

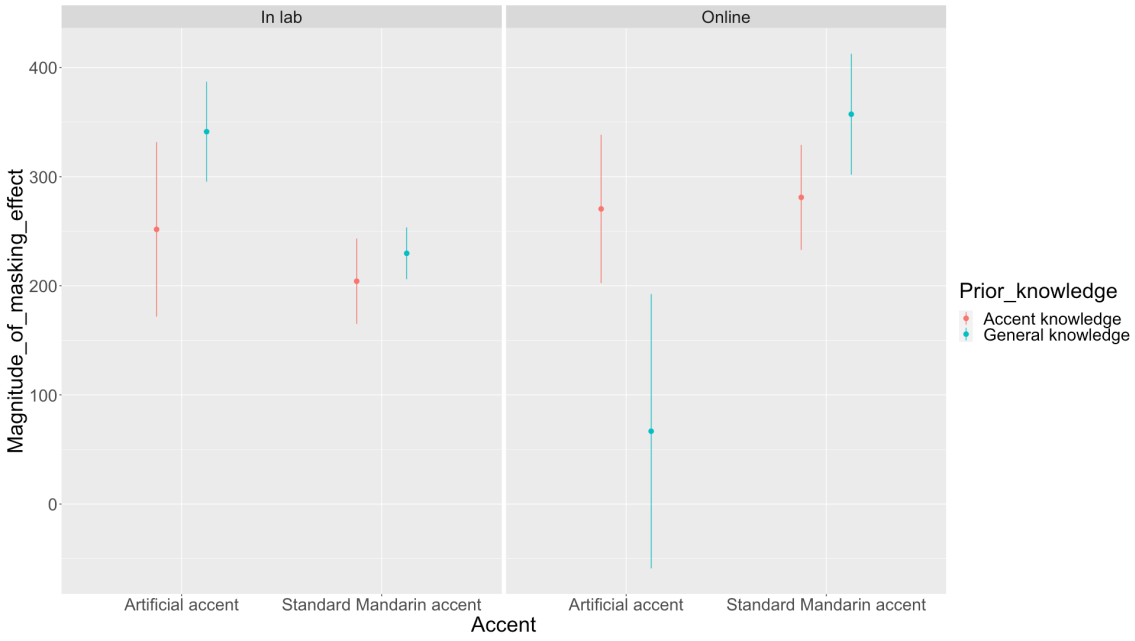


Figure 4.16: Experiment 2. Mandarin pseudo-word stimuli. Aggregated by-item. Masking effect of English listeners in milliseconds (ms) on y-axis. Conditions of *Accent* (Artificial accent/Standard Mandarin accent) are on x-axis. Fill color indicates the conditions of *Prior knowledge*(Accent knowledge/General knowledge). The facets indicate the conditions of *Modality* (In lab/Online).

Effect	DFn	DFd	F	p
pk	1	16	0.02	0.89
modality	1	16	0.03	0.86
accent	1	16	0.87	0.36
pk:modality	1	16	1.31	0.27
pk:accent	1	16	1.91	0.19
modality:accent	1	16	5.56	< 0.05
pk:modality:accent	1	16	3.24	0.09

Table 4.9: Experiment 2. Mandarin pseudo-word stimuli. Aggregated by-item. ANOVA of masking effect for English listeners. Prior knowledge is referred to as *pk*.

By-subject analysis

Masking effect of English listeners, along with the corresponding statistical analyses and variable interactions conducted through ANOVA, are depicted in Figure 4.17 and summarized in Table 4.10.

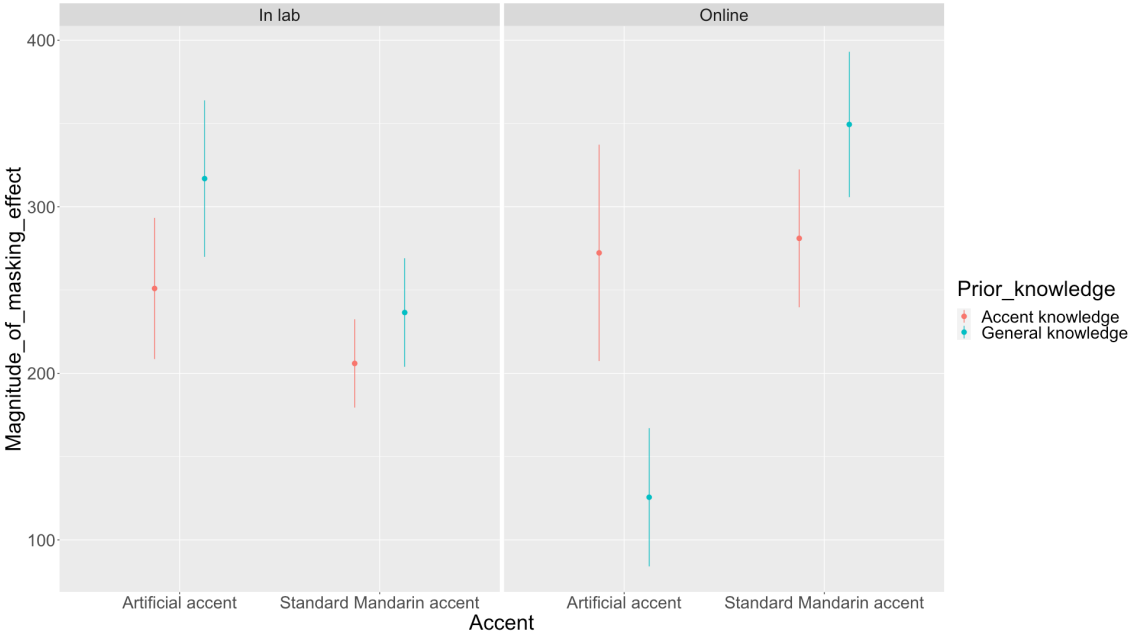


Figure 4.17: Experiment 2. Mandarin pseudo-word stimuli. Aggregated by-subject. Masking effect of English listeners in milliseconds (ms) on y-axis. Conditions of *Accent* (Artificial accent/Standard Mandarin accent) are on x-axis. Fill color indicates the conditions of *Prior knowledge*(Accent knowledge/General knowledge). The facets indicate the conditions of *Modality* (In lab/Online).

Effect	DFn	DFd	F	p
pk	1	52	0.0002	0.99
modality	1	52	0.09	0.77
accent	1	52	1.37	0.25
pk:modality	1	52	1.56	0.22
pk:accent	1	52	3.78	0.06
modality:accent	1	52	8.70	< 0.01
pk:modality:accent	1	52	4.51	< 0.05

Table 4.10: Experiment 2. Mandarin pseudo-word stimuli. Aggregated by-subject. ANOVA of masking effect for English listeners. Prior knowledge is referred to as *pk*.

Accent as bottom-up information did not show main effect (F(1, 52) = 1.37, p = 0.25). However, Accent has an interaction with Modality (F(1, 52) = 8.70, $p < 0.01$). As showed in Figure 4.17, there is a decreased masking effect in the online condition. The follow-up t-test (paired t-test, *Accent* as within-subject variable) confirmed that when the experiment was conducted online, Accent was found to have a significant effect (t(27) = -2.64, $p < 0.05$) in reducing masking effect. Thus, English listeners, as non-native listeners of Mandarin pseudo-words, showed a more significant effect of Accent interacting with Modality (in an online setting) than Mandarin listeners, who showed neither a main effect of Accent (F(1, 76) = 1.77, p = 0.22) nor an interaction effect with Modality (F(1, 76) = 0.60, p = 0.44).

Figure 4.18 below showed the interaction between Accent and Modality, on the RTs to phoneme and syllable targets. As presented in Figure 4.18, in the online condition, when artificial accent was applied, RTs to syllables increased more than RTs to phonemes (left column, blue point ranges changed more drastically than orange point ranges), resulting in a smaller masking effect compared to when standard Mandarin accent was applied.

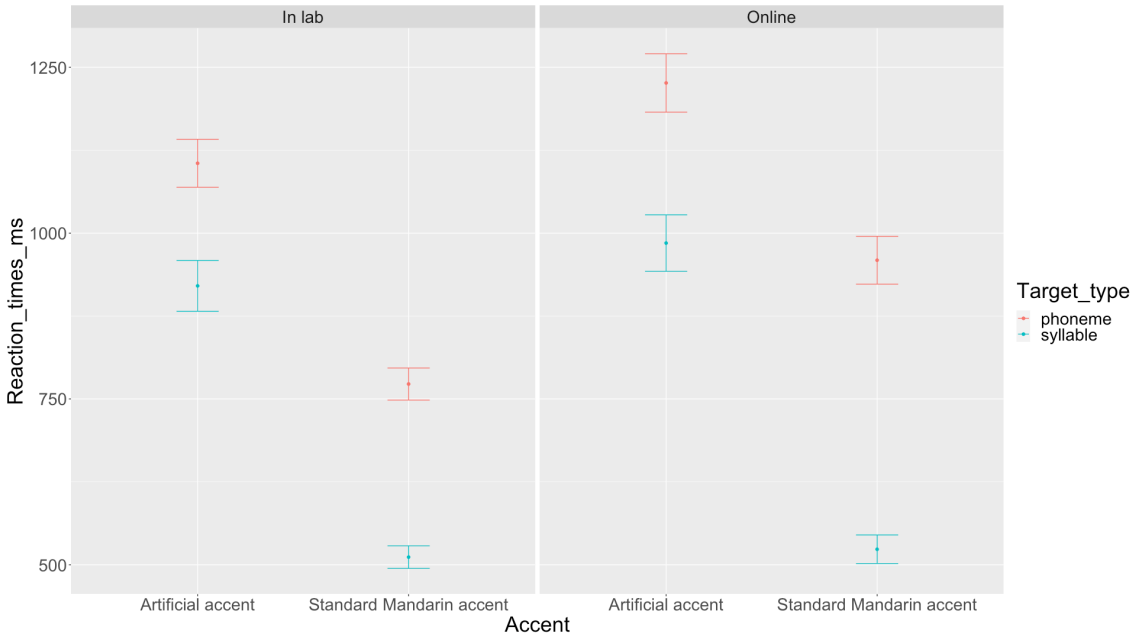


Figure 4.18: Experiment 2. Mandarin pseudo-word stimuli. RTs to phoneme and syllable targets for English listeners in milliseconds (ms) on y-axis, showing interaction between *Accent* and *Modality*.

Further, prior knowledge did not show a main effect (F(1, 52) = 0.0002, p = 0.99). The

main effect of Modality was also insignificant ($F(1, 52) = 0.09$, $p = 0.77$). However, there was a significant three-way interaction between Modality, Prior Knowledge, and Accent ($F(1, 52) = 4.51$, $p < 0.05$). As presented in Figure 4.17, in the online condition, the artificial accent significantly decreased the masking effect with general knowledge (showed by the bottom right blue point range in Figure 4.17). Figure 4.19 below displays this three-way interaction resulting from the RTs to phonemes and syllables. At the bottom right in the Figure 4.19, there is a significant increase in RTs to both phonemes and syllables, but with a greater increase for syllables than phonemes when the artificial accent is applied with general knowledge in the online condition. This leads to the decreased masking effect showed in Figure 4.17.

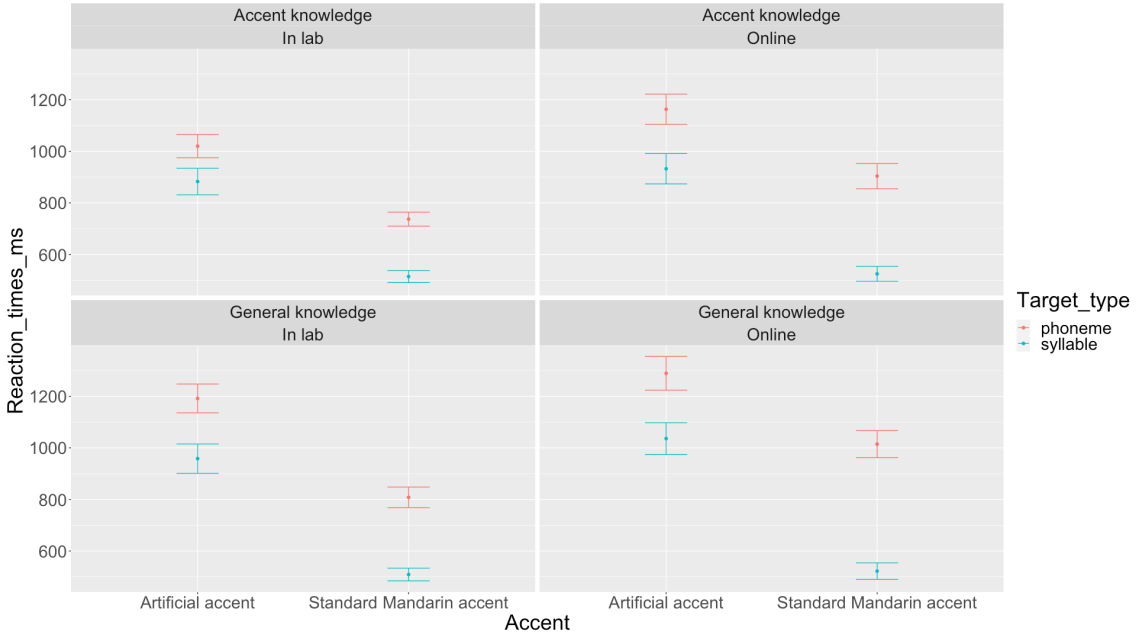


Figure 4.19: Experiment 2. Mandarin pseudo-word stimuli. RTs to phoneme and syllable targets for English listeners in milliseconds (ms) on y-axis, showing interaction between *Accent*, *Prior knowledge* and *Modality*.

A typical masking effect was also observed in English listeners across all conditions, as showed in Figure 4.17, indicating that even with the use of an artificial accent as bottom-up information and prior knowledge as top-down information to improve the perception of phoneme targets, listeners responded more rapidly to syllable targets than to phoneme targets. Figure 4.20 below illustrates the RTs of English listeners to phonemes and syllable. Overall, Accent and Prior knowledge did not exhibit a main effect on masking but showed three-way interaction with Modality.

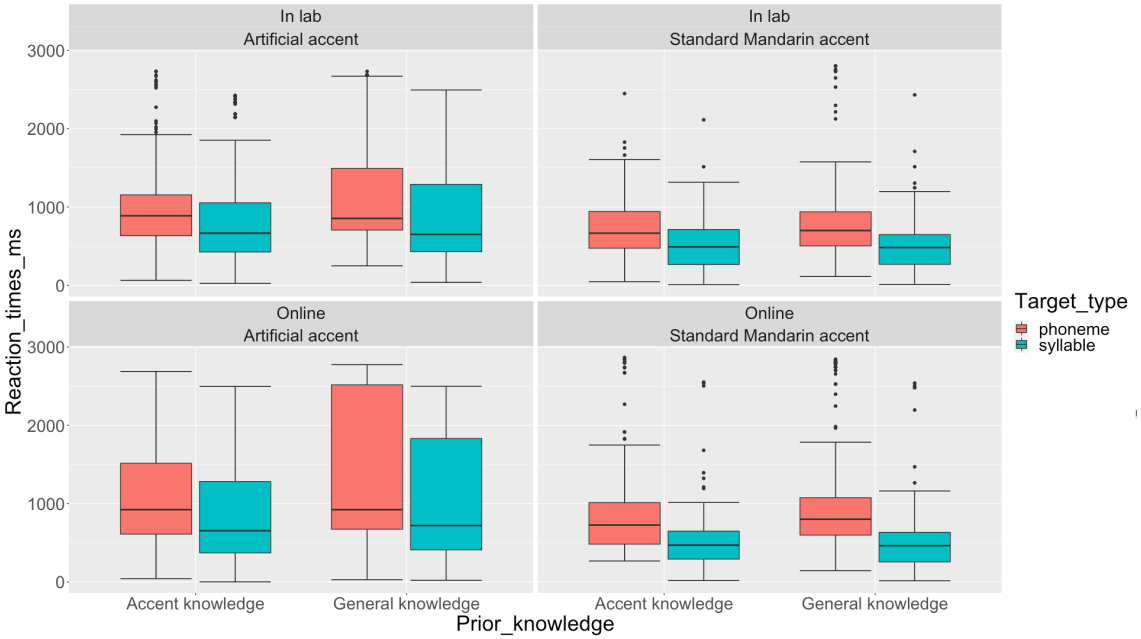


Figure 4.20: Experiment 2. Mandarin pseudo-word stimuli. RTs to phonemes and syllables for English listeners in milliseconds (ms) on the y-axis and the conditions of *Prior knowledge* (Accent knowledge/General knowledge) on the x-axis. Fill color indicates *Target type* (Phoneme/Syllable). Facets indicate the conditions of *Accent* (Artificial accent/Standard Mandarin accent) and conditions of *Modality* (In lab/Online).

4.3.3 Summary

In Experiment 2, under the Mandarin pseudo-word context, both groups of listeners consistently showed a typical masking effect by reacting faster to syllables than to phonemes. This is similar to the pattern demonstrated in Experiment 1. Similarly to Experiment 1, a by-item analysis also showed few significant effects. This could suggest that the variability between items in Mandarin stimuli is low compared to the variability between subjects. Differences in participant responses are less likely to be driven by the specific characteristics of the individual item.

Unlike Experiment 1, a by-subject analysis showed that under Mandarin pseudo-word stimuli, bottom-up information had a more substantial impact on drawing listeners' attention to syllables than prior knowledge as top-down information, whose impact was found to be little in Experiment 2. Also differing from Experiment 1, in Experiment 2, English listeners performed differently from Mandarin listeners. Mandarin listeners showed no main effects or interactions among Accent, Prior Knowledge, and Modality.

In contrast, English listeners' identification of syllables showed significant sensitivity to bottom-up information. When interacting with Modality, the artificial accent significantly reduced the masking effect under the online condition, resulting from a greater increase in RTs to syllables than to phonemes. This implies that when the experiment was conducted in participants' own environments, non-native listeners found the artificial accent more challenging than in the lab setting. This could be attributed to differences in environmental conditions and levels of attention. Furthermore, when interacting with both Modality and Prior Knowledge, the artificial accent, combined with general knowledge under the online condition, substantially reduced the masking effect, again due to a greater increase in RTs to syllables than to phonemes.

Therefore, English listeners, as non-native listeners in the Mandarin pseudo-word context, showed significant fluctuations in the masking effect, indicating greater perceptual flexibility compared to Mandarin listeners. English listeners demonstrated more focus and sensitivity to syllables when dealing with the artificial accent as bottom-up information.

In other words, the accent, as bottom-up information, disrupted English listeners' syllable processing more than their phoneme processing. This pattern is similar to how English listeners behaved in Experiment 1 with English pseudo-word stimuli.

4.4 Discussion

The findings regarding the different patterns of perceiving perceptual units for native and non-native listeners confirm some, but not all, of the predictions. As outlined in Section 3.1 *Experiment overview*, the current monitoring experiment aims to reveal perceptual flexibility in speech unit perception by measuring the masking effect. According to the concept of perceptual flexibility, listeners should exhibit a significant change in the magnitude of the masking effect when bottom-up and top-down information are manipulated. This implies that listeners' attention can be flexibly directed to a specific level of unit, thereby influencing the RTs to phonemes and syllables.

Native listeners were expected to show higher perceptual flexibility, which was anticipated to result in greater changes in RTs to phonemes than to syllables. This expectation is based on the fact that both bottom-up and top-down information were applied at the phoneme level, and native listeners have been showed to be able to direct their attention to phoneme-level units (Goldinger & Azuma, 2003). However, the results from Experiments 1 and 2 showed that the fluctuation in the masking effect was driven more by the listeners' syllable RTs fluctuating than by their phoneme RTs.

In general, it is not surprising to observe that in Experiments 1 and 2, the artificial accent, as bottom-up information, increased RTs to both phonemes and syllables, while prior knowledge of the artificial accent reduced the difficulty of dealing with it. This aligns with the understanding that the methods used in the current monitoring experiment differ from those in Goldinger and Azuma's study (2003). The artificial accent in this study, similar to distorted versions of phonemes, replicated unfamiliar accents encountered in daily conversations, unlike the acoustic cues enhancement in Goldinger and Azuma's research. To boost phonemes detection, their modifications resulted from encouraging speakers to produce their phonemes more clearly, which resulted in enhancements such as lengthened

fricatives and pronounced stop releases, rather than distortions. These enhancements provided clear cues, making phonemes more discernible than their carrier syllables, which led to a reversed masking effect resulting from decreased RTs for phonemes.

In the following sections, the discussion is from two angles. First, there is a general observation that changes in bottom-up input could disrupt syllable processing more than phoneme processing, suggesting that syllables attract more focus and attention than phonemes when processed under accented stimuli, particularly for English listeners. Second, the generally comparable perceptual flexibility found in native and non-native listeners is in line with exemplar models' assumptions about speech unit representation.

4.4.1 Phonemes as more robust unit than syllables

In both Experiments 1 and 2, syllable processing was more disrupted than phoneme processing, especially by the artificial accent applied in stimuli as bottom-up information. Initially, listeners' attention and focus were expected to be directed to the phoneme level by both top-down and bottom-up information, particularly for native listeners. However, with changes in bottom-up input, processing difficulty increased more for syllables than for phonemes. As a result, generally, both native and non-native listeners expended more time processing syllables than phonemes. Consequently, under the influence of both bottom-up and top-down information, syllable processing may attract greater sensitivity and focus from listeners than phoneme processing.

Moreover, the artificial accent, as bottom-up information, generally had a more significant impact on RTs to targets than prior knowledge as top-down information. Particularly in the context of English pseudo-words, both native and non-native listeners exhibited significant fluctuations in the magnitude of the masking effect when the artificial accent was applied, with a more pronounced delay in RTs for syllables compared to phonemes. This suggests that the lesser disruption to phoneme processing indicates that the phoneme may be a more fundamental and robust unit during speech processing. In contrast, larger units like syllables are more easily affected by accented speech or distorted sounds, a finding supported by the current monitoring experiment.

4.4.2 MINERVA 2 highlighting potential differences between native and non-native speech perception

During speech unit perception, native and non-native listeners displayed minimal differences in perceptual flexibility. Under English pseudo-word context, English listeners as native listeners showed a more substantial changes in the masking effect under the bottom-up information manipulation compared to Mandarin listeners as non-native listeners, caused by a more drastic increase in RTs to syllables than to phonemes. This suggests that, under the English pseudo-word context, native listeners showed greater sensitivity to the artificial accent, which makes the syllables even harder for them to identify.

Results from Experiments 1 and 2 aligned with how MINERVA 2 works outlined in the Section 2.4.1.2 *Implementation of mathematics in MINERVA 2*, where native listeners are thought to have more extensive feature vectors than non-native listeners. Consequently, the feature vector of the accented speech unit is more likely to contradict the feature vector of the traces, containing more inhibitory features *(-1)* corresponding to the original positive features *(1)* of the non-accented phoneme in the traces. The results of Experiment 1, showing that bottom-up information disrupted English listeners' syllable processing more significantly than it did for Mandarin listeners, validate the MINERVA 2 unit representation, which characterizes that, compared to native listeners, non-native listeners are less "sensitive" to the artificial accent as bottom-up information during speech perception.

However, it was unexpected to observe in Experiment 2, that under the Mandarin pseudo-word context, English listeners as non-native listeners exhibited more substantial changes in the masking effect under bottom-up information manipulation in an online setting compared to Mandarin listeners as native listeners. This result could also be explained by MINERVA 2, where non-native listeners may have inaccurate feature representations in their feature vectors, leading to more mismatches when matching input with stored traces, compared to native listeners. Therefore, this situation suggests that, compared to native listeners, non-native listeners are more "sensitive" to the artificial accent as bottom-up information during speech perception.

The opposite behavior of non-native listeners may be due to the proficiency levels of the non-native participants recruited in Experiments 1 and 2, as discussed in Chapter 8 *Limitations*. The non-native listeners in Experiment 2, whose L1 is English and L2 is Mandarin, may have had lower proficiency compared to the non-native listeners in Experiment 1, whose L1 is Mandarin and L2 is English. This could be largely because the English native listeners learning Mandarin as their L2 were recruited from an English-speaking environment, but not from Mandarin-speaking environment. Consequently, non-native listeners in Experiment 2, with lower L2 proficiency compared to those in Experiment 1, may still have faced greater challenges in storing accurate feature vectors. This could explain why, compared to native listeners, non-native listeners in Experiment 2 appeared to be more "sensitive" to the artificial accent as bottom-up information.

4.4.3 Post-hoc examination of data

To ensure the data is scientifically justified, the following analysis examines the data from Experiments 1 and 2 for masking effect after merging the online and in-lab conditions and removing the Modality factor.

4.4.3.1 Experiment 1

By-item analysis for English listeners

When the data were aggregated by item, English listeners showed neither a significant main effect of Accent nor Prior Knowledge, nor significant interactions between them.

Effect	DFn	DFd	F	p
pk	1	8	0.18	0.68
accent	1	8	0.72	0.42
pk:accent	1	8	0.01	0.92

Table 4.11: Post-hoc examination for Experiment 1. English pseudo-word stimuli. Aggregated by-item. ANOVA of masking effect for English listeners. Prior knowledge is referred to as *pk*.

By-subject analysis for English listeners

When the data were aggregated by participants, as shown in Figure 4.21 and Table 4.12, the results were similar to the previous analysis. The main effect of Accent was significant ($F(1, 78) = 8.89$, $p < 0.01$), while the main effect of Prior Knowledge was marginally significant ($F(1, 78) = 4.07$, $p = 0.05$). As showed in Figure 4.21, there is a significant decrease in the masking effect when the artificial accent was applied. Therefore, consistent with the previous analysis, English listeners were more strongly influenced by bottom-up information than by top-down information. Further, there was no interaction between the effects of Prior Knowledge and Accent ($F(1, 78) = 0.30$, $p = 0.59$).

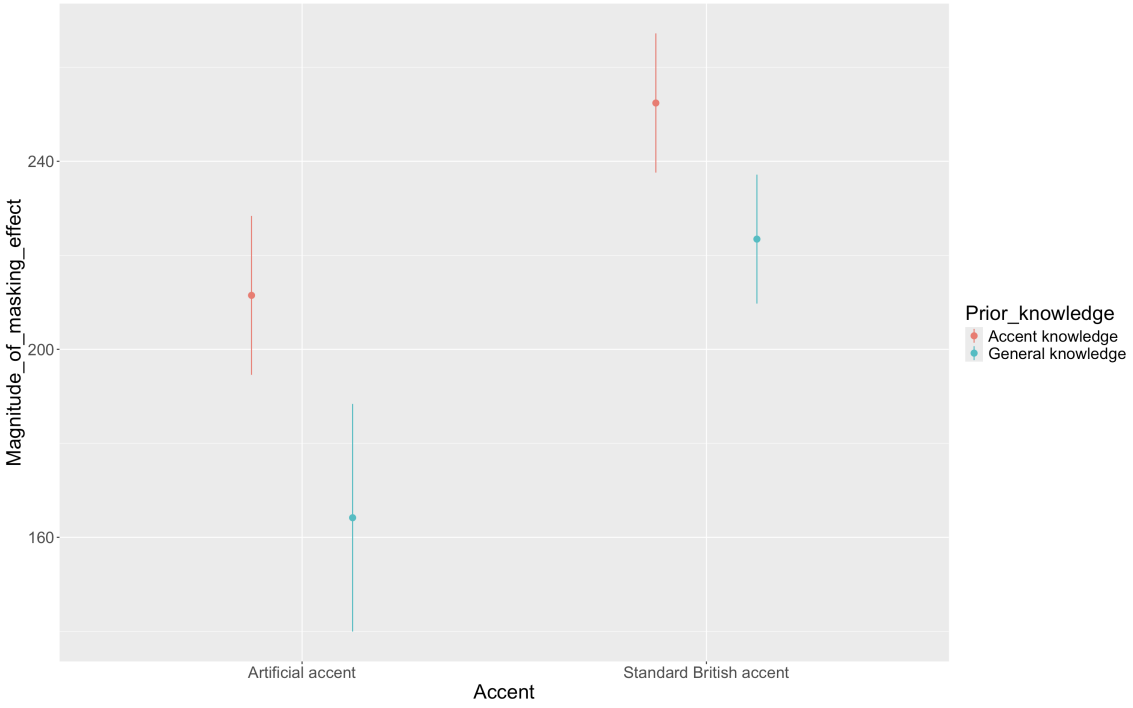


Figure 4.21: Post-hoc examination for Experiment 1. English pseudo-word stimuli. Aggregated by-subject. Masking effect for English listeners in milliseconds (ms) on y-axis. Conditions of *Accent* (Artificial accent/Standard British accent) are on x-axis. Fill color indicates the conditions of *Prior knowledge*(Accent knowledge/General knowledge).

Effect	DFn	DFd	F	p
pk	1	78	4.07	0.05
accent	1	78	8.89	< 0.01
pk:accent	1	78	0.30	0.59

Table 4.12: Post-hoc examination for Experiment 1. English pseudo-word stimuli. Aggregated by subject. ANOVA of masking effect for English listeners. Prior knowledge is referred to as *pk*.

As showed in Figure 4.22 below, the artificial accent slowed down the RTs to both types of targets, with RTs increasing more for syllables than for phonemes, leading to a decreased masking effect showed in Figure 4.21.

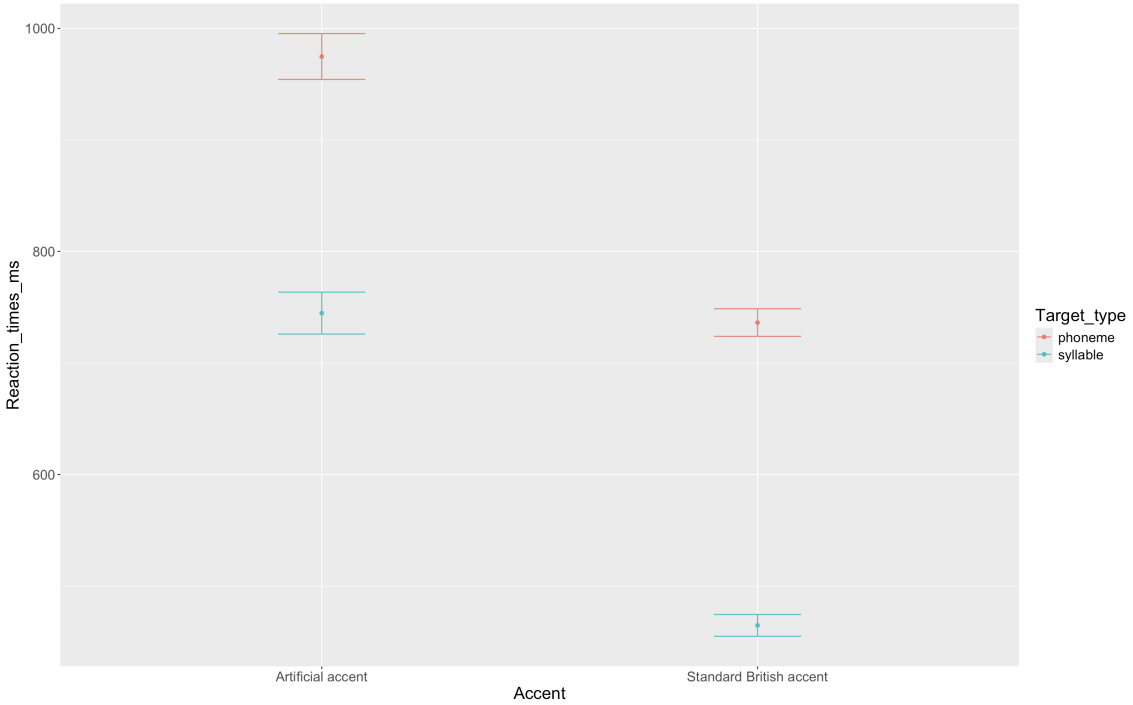


Figure 4.22: Post-hoc examination for Experiment 1. English pseudo-word stimuli. RTs to phoneme and syllable targets for English listeners in milliseconds (ms) on y-axis, showing effect of *Accent*.

By-item analysis for Mandarin listeners

When the data were aggregated by item, Mandarin listeners showed neither significant main effects for Prior Knowledge and Accent nor any interaction.

Effect	DFn	DFd	F	p
pk	1	8	0.40	0.55
accent	1	8	0.30	0.60
pk:accent	1	8	0.40	0.54

Table 4.13: Post-hoc examination for Experiment 1. English pseudo-word stimuli. Aggregated by-item. ANOVA of masking effect for Mandarin listeners. Prior knowledge is referred to as *pk*.

By-subject analysis for Mandarin listeners

When the data were aggregated by participants, the masking effect for Mandarin listeners is illustrated in Figure 4.23 and Table 4.14. Consistent with the previous analysis, the effects of Prior Knowledge (F(1, 78) = 4.57, $p < 0.05$) and Accent (F(1, 78) = 4.72, $p < 0.05$) are significant, with no interaction between them (F(1, 78) = 2.28, p = 0.13). As shown in Figure 4.23, there is a decrease in the masking effect when an artificial accent is applied, and an increase in the masking effect when prior knowledge is provided.

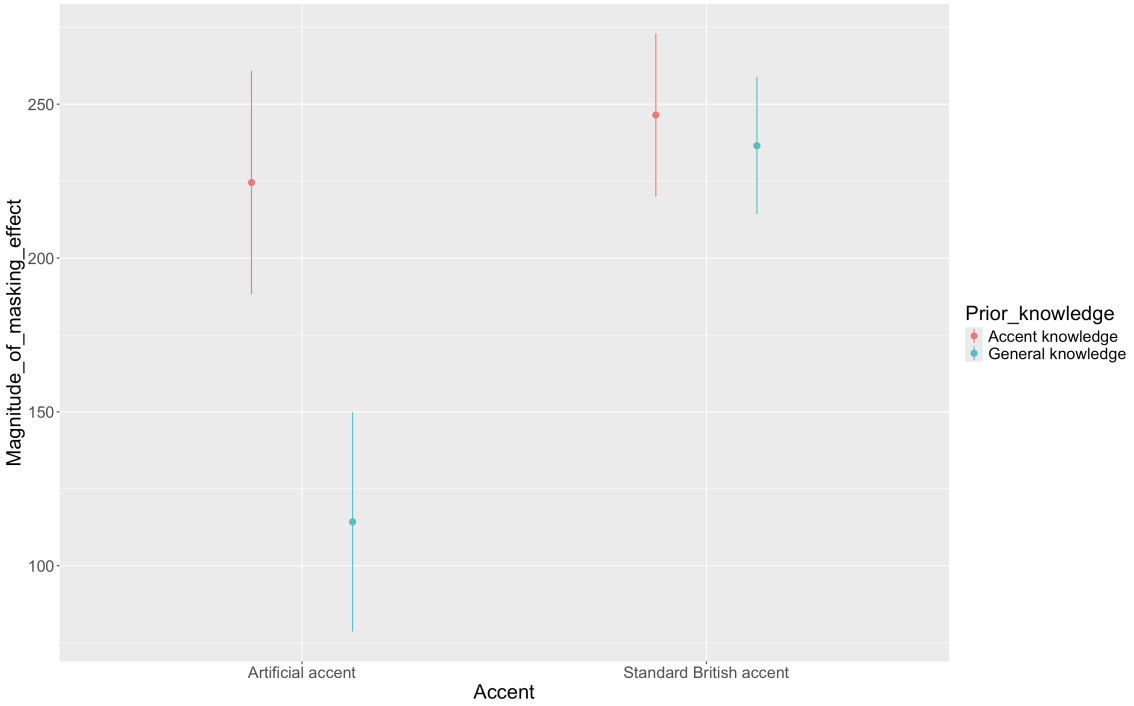


Figure 4.23: Post-hoc examination for Experiment 1. English pseudo-word stimuli. Aggregated by-subject. Masking effect of Mandarin listeners in milliseconds (ms) on y-axis. Conditions of *Accent* (Artificial accent/Standard British accent) are on x-axis. Fill color indicates the conditions of *Prior knowledge*(Accent knowledge/General knowledge).

Effect	DFn	DFd	F	p
pk	1	78	4.57	$<$0.05
accent	1	78	4.72	$<$0.05
pk:accent	1	78	2.28	0.13

Table 4.14: Post-hoc examination for Experiment 1. English pseudo-word stimuli. Aggregated by-subject. ANOVA of masking effect for Mandarin listeners. Prior knowledge is referred to as *pk*.

A similar pattern was found, as in the previous analysis, as shown in Figure 4.24. The application of prior knowledge decreased RTs for both phonemes and syllables, with greater facilitation of syllable detection than for phonemes, resulting in an increase in the masking effect showed in Figure 4.23.

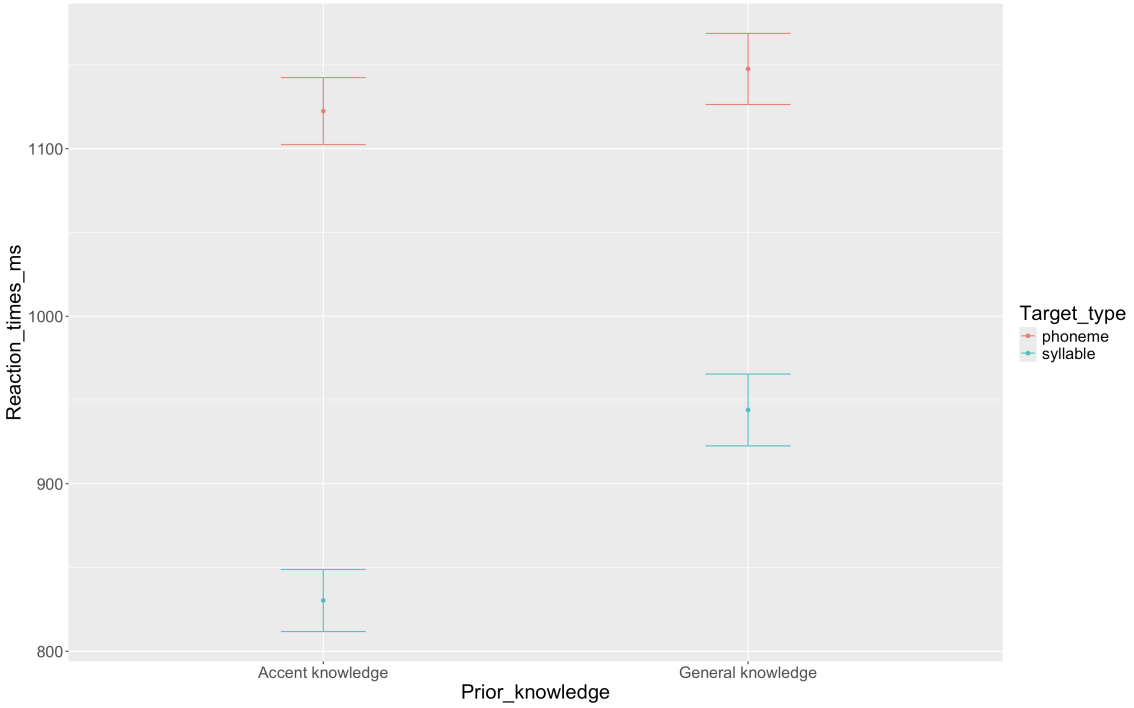


Figure 4.24: Post-hoc examination for Experiment 1. English pseudo-word stimuli. RTs to phoneme and syllable targets for Mandarin listeners in milliseconds (ms) on y-axis, showing effect of *Prior knowledge*.

Further, as shown in Figure 4.25, the application of an artificial accent slowed RTs for both phoneme and syllable targets, with a greater increase in RTs for syllables than for phonemes, leading to a decrease in the masking effect presented by Figure 4.23.

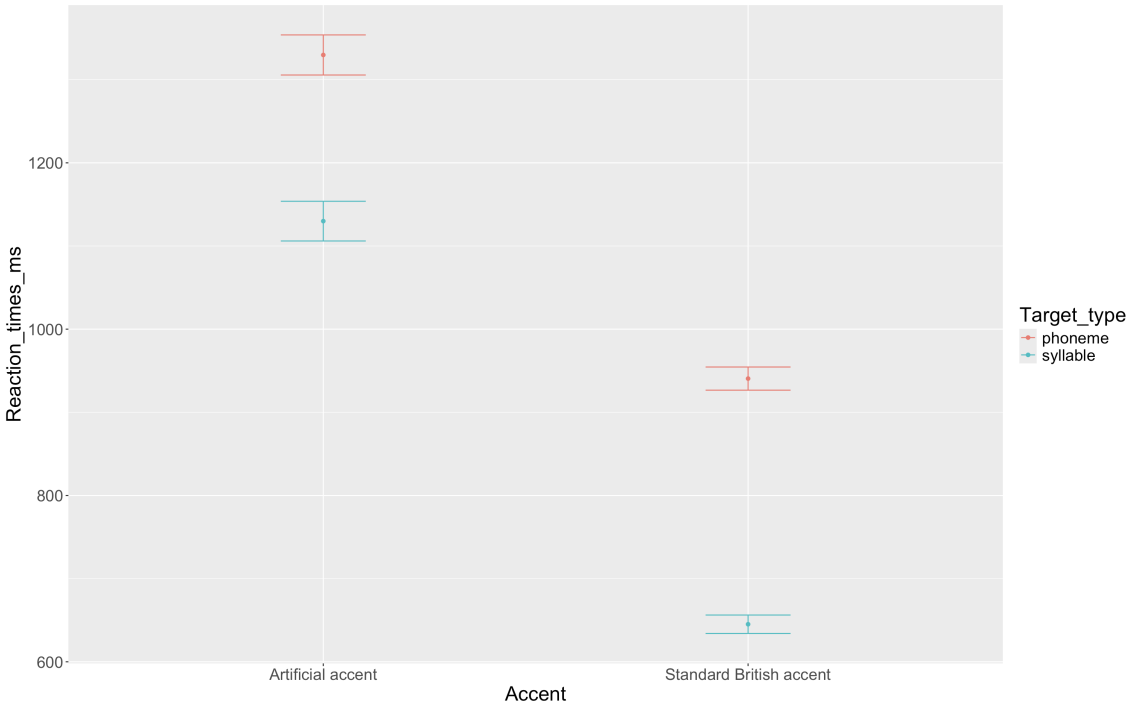


Figure 4.25: Post-hoc examination for Experiment 1. English pseudo-word stimuli. RTs to phoneme and syllable targets for Mandarin listeners in milliseconds (ms) on y-axis, showing effect of *Accent*.

4.4.3.2 Experiment 2

By-item analysis for Mandarin listeners

The by-item analysis showed that there were no significant main effects for Prior Knowledge and Accent, nor any interaction between them.

Effect	DFn	DFd	F	p
pk	1	8	0.07	0.79
accent	1	8	2.23	0.17
pk:accent	1	8	0.02	0.90

Table 4.15: Post-hoc examination for Experiment 2. Mandarin pseudo-word stimuli. Aggregated by-item. ANOVA of masking effect for Mandarin listeners. Prior knowledge is referred to as *pk*.

By-subject analysis for Mandarin listeners

Showed by Figure 4.16, consistent with the analysis of Mandarin listeners in Section 4.3, there were no main effects for Prior Knowledge and Accent, nor any interaction between them. This suggests stable processing of Mandarin phonemes and syllables for Mandarin listeners under the influence of prior knowledge and an artificial accent.

Effect	DFn	DFd	F	p
pk	1	78	0.01	0.92
accent	1	78	2.77	0.09
pk:accent	1	78	0.98	0.32

Table 4.16: Post-hoc examination for Experiment 2. Mandarin pseudo-word stimuli. Aggregated by subject. ANOVA of masking effect for Mandarin listeners. Prior knowledge is referred to as *pk*.

By-item analysis for English listeners

When the data were aggregated by item, no significant main effects or interaction between Prior Knowledge and Accent were found among English listeners.

Effect	DFn	DFd	F	p
pk	1	8	0.004	0.95
accent	1	8	0.44	0.52
pk:accent	1	8	1.24	0.30

Table 4.17: Post-hoc examination for Experiment 2. Mandarin pseudo-word stimuli. Aggregated by-item. ANOVA of masking effect for English listeners. Prior knowledge is referred to as *pk*.

By-subject analysis for English listeners

As shown in Table 4.18, there are no significant main effects for Prior Knowledge and Accent, nor any interaction between them. This suggests that English listeners, when processing Mandarin pseudo-words, like their native Mandarin counterparts, exhibit stable processing of phoneme and syllable targets.

Effect	DFn	DFd	F	p
pk	1	54	0.02	0.89
accent	1	54	0.73	0.40
pk:accent	1	54	2.05	0.16

Table 4.18: Post-hoc examination for Experiment 2. Mandarin pseudo-word stimuli. Aggregated by-subject. ANOVA of masking effect for English listeners. Prior knowledge is referred to as *pk*.

4.4.3.3 Discussion

For this post-hoc examination, when the data is aggregated by item, consistent with the analysis in Sections 4.2 and 4.3, the post-hoc examination revealed no significant main effects or interactions, suggesting low item variability. For the by-subject analysis, the post-hoc examination of the data for Experiment 1 showed results consistent with the

analysis in Section 4.2. For English listeners, significant main effects were found for both prior knowledge and Accent, with the artificial accent as bottom-up information having a greater influence on listeners. Compared to English listeners, Mandarin listeners showed comparable significant main effects for both prior knowledge and artificial accent, but, similar to their English counterparts, exhibited no interaction.

For Experiment 2, Mandarin listeners exhibited a consistent behavior pattern, as described in Section 4.3. For English listeners, in contrast to the analysis in Section 4.3, which showed a significant effect of artificial accent under the online condition, removing the Modality factor revealed that neither Prior Knowledge nor Accent had significant main effects, nor was there any interaction between them. This suggests that English listeners exhibited a behavior pattern similar to Mandarin listeners when processing Mandarin phoneme and syllable targets. Notably, both native and non-native listeners exhibited stable syllable processing, which was insensitive to prior knowledge and artificial accent. This could further suggest that Mandarin and English exhibit language-specific differences in phoneme and syllable processing.

4.4.4 Conclusion

In conclusion, differences in perceptual flexibility between native and non-native listeners during speech unit perception were generally minimal. Under the English pseudo-word context, native listeners exhibited higher perceptual flexibility by shifting their attention to syllables under the manipulation of bottom-up and top-down information. However, non-native listeners also showed comparable flexibility in shifting focus to syllables and experienced substantial fluctuations in the masking effect, similar to native listeners.

Under the Mandarin pseudo-word context, non-native listeners showed higher and more substantial changes in the masking effect than native listeners. Mandarin listeners in the Mandarin pseudo-word context were very stable in perceiving speech units and consistently demonstrated a stable masking effect, without showing a significant processing difficulty for syllables over phonemes. This could suggest that the impact of phoneme distortion on syllable processing is language-specific. Phoneme-level manipulations, such as

the retroflex /ɖʂ/, ejective /k'/ and palatal /ɲ/, may significantly disrupt syllable processing in English but have a less pronounced effect in Mandarin. For example, syllable-initial ejectives introduce a "disruption" (a mini glottal stop) into the syllable, which may not significantly affect phoneme monitoring since the initial sound is still /k/-like. However, it could impact syllable identification. For English listeners, a palatal nasal /ɲ/ may make a syllable sound like an extra segment has been added, as /n/ followed by /j/. In English, listeners might be robust to this in phoneme processing, it could disrupt syllable processing. In contrast, this phoneme distortion may have little effect on syllable-level processing in Mandarin. This finding leads to further investigation into how phoneme-level manipulations influence syllable processing in Mandarin (as discussed in Chapter 8, Section 8.3 *Conclusion and future directions*).

The perceptual flexibility in perceiving speech units, as revealed by the masking effect for both native and non-native listeners, further sheds light on and complements exemplar theories. First, for unit retrieval, phoneme processing to be less disrupted than syllable processing, suggesting the fact that phonemes could be the primary perceptual unit for both native and non-native listeners, especially under English context. Second, regarding unit representation, proficient non-native listeners may exhibit less sensitivity to unfamiliar accents compared to native listeners due to less detailed feature vectors. However, lower L2 proficiency may result in less accurate feature vectors in the L2. Non-native listeners with lower L2 proficiency might demonstrate greater sensitivity to unfamiliar accents compared native listeners.

Chapter 5

Perceptual flexibility of sentence processing: cue utilization and integration

5.1 Theories explaining the differences in utilizing cues during native and non-native speech processing

The previous chapters have highlighted the differences in perceptual flexibility for native and non-native listeners at level of speech unit perception. This chapter shifts the focus from phonetic unit perception to sentence processing. By exploring how listeners' attention is directed by multiple language cues during real-time sentence processing, this chapter examines perceptual flexibility for listeners from a different perspective: cue utilization and integration.

Native listeners have a higher inclination towards effortlessly and automatically employing and integrating a range of cues (Cutler, 2012; K. Ito & Speer, 2008; Patterson et al., 2017; Pisoni, 1981). Conversely, non-native listeners tend to show less preference for syntactic and prosodic cues (Akker & Cutler, 2003; Clahsen & Felser, 2006a; Hopp, 2010; Jiang, 2007; McDonald, 2006) and often face difficulties in the integration of cues across different domains (Patterson et al., 2017; Perdomo & Kaan, 2021). Therefore, mirroring the benefits observed in native sentence processing, this investigation aims to uncover whether such processes of integrating cues pose greater challenges for non-native listeners and, if so, to understand the underlying reasons and theories. Specifically, it examines which aspect—utilization, integration, or both—presents the significant difficulty

for non-native listeners.

Utilization and integration of cues emerge as two critical aspects during real-time sentence processing. Utilization and integration as two aspects is closely tied to the hierarchical nature of linguistic information processing. Sentence processing is described as a process mapping from "lower linguistic level" to "high linguistic level" from the sound signals, phonology, syntax, semantics, then to the message (Pickering & Garrod, 2013). Therefore, language processing relies on the interplay between acoustic-phonetic elements and linguistic-semantic aspects (Ezzatian et al., 2010). Proficiency in listening requires the ability to simultaneously engage with two sources of information and integrate them in real-time. This process involves the transformation of low-level characteristics from the bottom-up perceptual input into sublexical or lexical representations. These representations are then combined to form syntactic and semantic structures, ultimately constructing of higher-level linguistic meaning. Higher-level processing encompasses the extraction of lexical, semantic, syntactic, and contextual information from the low-level signal, facilitating the derivation of meaning from the incoming stream of speech (Brown & Kuperberg, 2015; Meyer, 2018; Norris, 1994). Therefore, the utilization and integration of cues relies on the hierarchical organization of linguistic information, with each level contributing to the comprehension of spoken language in a coordinated manner.

There are clear advantages in both the utilization and integration of cues in native sentence processing, where native listeners inherently possess the ability to flexibly and adeptly employ and combine these cues. I use the term *cues utilization* to refer to the listener's ability to recognize and use various acoustic signals present in speech to understand the message. In native speech processing, listeners weave together prosodic signals, such as contrastive pitch accents on nouns, with the semantic content of words and syntactic signals. For example, in Perdomo and Kaan's study (2021), they used sentences such as *We ate Angela's cake but saved BENjamin's cake in the fridge*, with the conjunction *but* and contrastive pitch accent on *Benjamin*. The result showed that when hearing the conjunction *but*, listeners predicted changes in narrative direction, resulting in more looks to the new object before hearing the second proper name. Then they shifted their gaze back to

the same object when they detected the contrastive pitch accent on *Benjamin*. Regarding listeners' capacity for utilizing diverse linguistic information, Cutler (2012) concluded that listeners are able to draw on and integrate a wide range of information: from tiny segmental and suprasegmental to word- and sentence-level information, for example, information extracted at phonemes, allophones, and syllables (Cutler, 2012; Pisoni, 2018). Prosodic, semantic, lexical-thematic, morphosyntactic (grammar), frequency, pragmatic, and contextual information are also important cues that can be immediately used and integrated to generate predictions in real-time comprehension for listeners (Cutler, 2012; Hopp, 2015; Huettig et al., 2011; Kamide, 2008; Pisoni, 2018).

Regarding the utilization of prosody, native listeners, for example, use F0 and stress for segmentation and to identify word-initial boundaries in continuous speech (Coughlin & Tremblay, 2012; Cutler & Clifton, 1984; Cutler & Norris, 1988). They also utilize prosody at the verb to predict the upcoming noun: focusing on the verb through prosody leads listeners to make rapid inferences about the intended referent even before the noun is fully processed (Kurumada et al., 2014). Specifically, Kurumada et al.'s study (2014) applied the contrastive pitch accent L+H* on the verb *look* in a sentence like *It looks like a zebra, but it is not*. Listeners listened to these sentences while viewing a set of contrasting pictures (e.g. a zebra (prototypical) vs. a zebra-like animal (okapi, non-prototypical)). Kurumada et al. expected the contrastive pitch accent on the noun to bias responses toward the more prototypical member of each pair (e.g. zebra), while the contrastive pitch accent on the verb to bias responses toward the less prototypical member (e.g. okapi). The results showed that native listeners can utilize contrastive pitch accents on verbs to facilitate early consideration of specific referents in visual contexts.

Extracting cues during speech perception and establishing a predictive process is not exclusive to native speech perception; it is also observable in non-native speech perception. While non-native listeners demonstrate the ability to identify, utilize, and integrate various cues in speech perception, they exhibit significant delays and reduced ability in employing and integrating specific cues effectively compared to native listeners. Clahsen and Felser (2006b) highlighted that in the real-time processing of non-native speech, listeners do not

depend as much on syntactic information as they do with their native language. Further, the processing of prosodic cues, which increases the cognitive load, may be abandoned by non-native listeners during real-time speech processing (Akker & Cutler, 2003). Additionally, non-native listeners may struggle with integrating different cues (Hopp, 2015), especially when syntactic and prosodic cues are involved (Sorace, 2011; Sorace & Filiaci, 2006; Takahashi et al., 2018).

One of the most noticeable aspects for listeners to process the non-native speech is the absence of *Automaticity* (Segalowitz, 2003). This *Automaticity* in native real-time speech processing refers to the ability to process speech quickly and efficiently without conscious effort or delay, which suggests that the speech information is processed in an effortless, fast, and accurate manner. Listeners in native speech processing have the advantage of processing with *automaticity* and rapid integration of multiple cues. According to Segalowitz (2003), compared to native listeners, non-native listeners may struggle with automaticity in several areas such as vocabulary recognition, morphosyntactic processing, phonological processing, and pragmatic understanding. For example, they might not instantly recognize words, requiring more effort to understand spoken language, and a lack of automaticity in grammar can hinder their ability to parse sentences quickly and understand complex structures during real-time speech processing.

L2 proficiency, which includes factors such as years of exposure to the L2 environment, is often considered the primary explanation for the differing performance observed in listeners during native and non-native real-time sentence processing (Chun & Kaan, 2019; Dijkgraaf et al., 2017; Kilman et al., 2014; S. Lee et al., 2022; Xue & Lee, 2014). However, theories of sentence processing suggest that, beyond the level of proficiency, various factors can contribute to differences in performance levels between native and non-native speech processing. These factors include the influence and transfer from the L1 and the cognitive load involved in processing L2, which may influence listeners' interpretation and utilization of cues and force listeners to disregard certain cues during real-time speech processing. Clahsen and Felser (2006b) highlighted that the influence of a speaker's L1 could potentially cause delays in the cognitive processing of morphosyntax in a non-

native language. Other studies have showed that when processing non-native cues, such as the categorization of VOT (Vowel Onset Time) for voiced and voiceless stop consonants, listeners are influenced by the VOT categorization from their L1 (Schwartzhaupt et al., 2015). Additionally, non-native listeners may be affected by their L1's phonotactic rules regarding word-initial clusters when perceiving spoken English words, leading them to interpret English words more likely to conform to the phonotactic rule of their L1 (Freeman et al., 2022).

Recall that this investigation explores how the utilization and integration of cues pose significant challenges for non-native listeners. Accordingly, this section explores various theories and hypotheses that aim to explain the performance disparities observed between native and non-native listeners, with a particular emphasis on their ability to utilize and integrate cues in language comprehension and processing. It reviews the *Shallow Structure Hypothesis*, *Interface Hypothesis*, *Cue Weighting Theory*, and other speculations derived from previous studies. These theories and speculations highlight the challenges non-native listeners face in effectively utilizing and integrating cues, thereby providing a comprehensive overview of the factors contributing to the observed discrepancies in language processing abilities for native and non-native listeners.

5.1.1 Shallow structure hypothesis

The *Shallow Structure Hypothesis* was introduced by Clahsen and Felser (2006a, 2006b). The Shallow Structure Hypothesis places a greater focus on the reduced reliance on syntactic cues during language processing, indicating a preference for surface-level cues over deeper linguistic structures. The *Shallow Structure Hypothesis* posits that non-native speakers engage in superficial parsing, meaning they form syntactic representations that lack the depth and detail characteristic of native speakers' parsing in both speech and reading (Clahsen & Felser, 2006b).

Findings from prior research were aligned with Shallow structure hypothesis. For example, Left Anterior Negativity (LAN) and a centro-parietal positive wave (P600) are typically associated with syntactic anomalies (Meltzer & Braun, 2013). However, Hahne

and Friederici's study (2001) found neither the LAN effect nor the P600 for L2 learners of German when encountering word category violations (e.g. noun versus verb, as in **The ice cream was in the eaten*). This suggests that L2 listeners might skip the stage where initial syntactic structures are formed based on word category information.

Additionally, Hopp's research (2015) demonstrated that non-native speakers struggle to combine case marking as a morphosyntactic indicator with lexical-semantic cues. They tended to solely depend on the lexical-semantic information, neglecting the morphosyntactic cues. In Hopp's study (2015), English native speakers learning German as L2 were tested through an eye-tracking experiment while listening to German stimulus sentences. Two types of German stimulus sentences were involved: SVO (subject-verb-object) sentence *Der Wolf tötet gleich den Hirsch* ("The_{NOM} wolf kills soon the_{ACC} deer. "The wolf will soon kill the deer"); and OVS (object-verb-subject) sentence *Den Wolf tötet gleich der Jäger* ("The_{ACC} wolf kills soon the_{NOM} hunter." "The hunter will soon kill the wolf"). Results showed that English native listeners predicted the upcoming second noun in the sentences based solely on verb semantics, disregarding case marking adjustments. Consequently, English native listeners in this study consistently anticipated the second noun to be the patient in the given context. In contrast, German native listeners successfully integrated both case and verb semantics. Even English native speakers with high proficiency in German did not exhibit the ability to adroitly integrate these two types of information together during non-native speech perception. This finding highlights the challenges non-native listeners face in using multiple linguistic cues simultaneously for language comprehension and their preference for the semantic cue compared to the morphosyntactic cue.

Therefore, the shallow structure hypothesis explains why L2 learners may not exhibit certain effects seen in native speakers, such as the use of structure-based parsing strategies or the ability to resolve syntactic ambiguities based on phrase structure. Instead, non-native listeners typically concentrate on integrating new information into a developing semantic representation, guided by lexical and semantic cues, rather than mapping the incoming information into a detailed syntactic structure.

Listeners or readers comprehending a non-native language tend to focus more on surface features than on the underlying grammatical structure (Clahsen & Felser, 2006a). When the shallow structure hypothesis is applied to the utilization of cues in speech perception, it suggests that listeners during the speech processing tend to minimize the processing of syntactic structures or give reliance on surface cues for comprehension. Non-native listeners may adopt shallow parsing strategies or inadequate syntactic processing. This results in even highly proficient non-native listeners relying more on lexical, semantic, and pragmatic information to construct a semantic or conceptual representation of the sentence.

5.1.2 Interface hypothesis

The *Interface Hypothesis* was proposed by Sorace and Filiaci (2011, 2006, 2011), suggesting that while learners can fully acquire syntactic properties in a L2, the integration of syntactic cues with other cues, essentially, interface between syntax and other cognitive domains, may not be fully attainable. The Interface hypothesis aligns with the Shallow structure hypothesis mentioned in the preceding section, both the Interface hypothesis and the Shallow structure hypothesis suggest that during non-native speech processing, listeners are more likely to disregard or unable to fully utilize syntactic structures than they would in native speech processing. However, different from the Shallow structure hypothesis which emphasizes the reduced reliance on the syntactic cues in language processing, the Interface Hypothesis highlights the integration of syntactic cues with other cues.

For the *Interface Hypothesis*, the syntax and other cognitive domains, such as pragmatics and semantics, may present residual effect of L1 even among very advanced or near-native speakers of a L2. Sorace and Filiaci (2006) used an example from a reading task to illustrate this challenge of integrating syntax and pragmatics for non-native readers. To be specific, they tested the near-native L2 Italian speakers and discovered an inappropriate extension of the scope of overt subject pronouns for the near-native L2 Italian speakers. For example, it is unusual for the Italian pronoun *lei* to occur at the sentence initial position, showed in example (1). Therefore, it is marked and unusual for the sentences such

as:

(1) *Perche lei non ha trovato un taxi.*
Because she couldn't find a taxi.

In example (1), the pronoun *lei* at the sentence-initial position is pragmatically determined. Although the sentence-initial position is not normal, native speakers can still readily interpret it because they are more attuned to these pragmatic cues. On the contrary, example (2) showed that the normal position of the pronoun lei should be:

(2) *La vecchietta saluta la ragazza quando lei attraversa la strada.*
The old woman greets the girl when she crosses the road.

Sorace and Filiaci used a Picture Verification Task in which participants were presented with sentences containing the pronoun *lei* in sentence-initial position, such as *Mentre lei si mette il cappotto, la mammai da un bacio alla figliak* (While she is wearing her coat, the mother kisses her daughter). Participants had to choose the picture that matched the meaning of the subordinate clause (e.g., a picture where the character 'Mother' performs both the action in the main clause and the one described in the subordinate clause). Results showed that non-native speakers not always interpret the pronoun *lei* in its sentence-initial position as readily as native speakers do. This suggests that even the highly proficient speakers have difficulty integrating the multiple types of information at the interfaces between syntax and pragmatics. However, studies (Sorace, 2011; Sorace & Filiaci, 2006) also indicated that the broad applicability of the interface hypothesis across various bilingual domains, suggesting its relevance to speech perception as well. Therefore, based on the reasoning behind the Interface hypothesis, non-native sentence processing faces challenges in integrating syntax with other cognitive domains, unlike the rapid integration of lexical, discourse-level, prosodic, and structural information observed in native real-time sentence processing.

Both the *Shallow Structure Hypothesis* and *Interface Hypothesis* align with prior research (Hopp, 2010, 2015; McDonald, 2006), highlighting the difficulties non-native listeners face in using the syntactic and morphosyntatic cues during the real-time speech processing.

5.1.3 Cue weighting theory

Similar to the concept of *attention weight* from the Generalized Context Model (GCM, see *Chapter 2*) (Nosofsky, 1992; Nosofsky, 1986, 1988), *Cue weighting theory* also highlights that listeners have their preferences and sensitivity to particular or multiple acoustic properties or dimensions during speech perception and sentence processing. However, unlike *attention weight* in the GCM, the *Cue weighting theory* emphasizes that listeners will transfer the native weighting or preference of specific cues to non-native sentence processing. To be specific, the *Cue weighting theory* (Tremblay et al., 2018) posits that, due to differences in how language cues are weighted across languages, listeners can transfer their native language's use of cues, or the native weighting of specific cues, to non-native sentence processing. This means that the cues utilized in one's native language can influence the utilization of the same cues in non-native sentence processing.

For example, in English, Dutch, and French, the fundamental frequency (F0) rise is employed as a cue to signal word-final boundaries. Tremblay et al. (2018) examined the use of F0 rise by English and Dutch listeners in their L2 French speech perception, particularly in the context of determining word-final boundaries. The results demonstrated that Dutch listeners used F0 rise earlier and to a greater extent than English listeners. This is attributed to the greater functional weight of the F0 rise in Dutch compared to English. Likewise, the studies (Ingvalson et al., 2012; Iverson et al., 2003) on Japanese listeners' perception of alveolar lateral approximants /l/ and alveolar approximants /ɹ/ reveal that native language experience can negatively impact the perception of non-native phonemes in adulthood. In the study of Iverson et al. (2003), to categorize the English /ɹ/-/l/, Japanese native listeners exhibit a greater sensitivity to the second-formant (F2) cue compared to the third-formant (F3) cue. This is because F2 cue is crucial for perceiving the Japanese liquid sound but is not relevant and useless for distinguishing /l/ and /ɹ/, which primarily

vary in F3 onset frequency. Therefore, compare to the F2 cue, F3 cue is the important cue to distinguish /l/ and /ɹ/ in English which vary only in F3 onset frequency. Consequently, the difficulty Japanese listeners face in distinguishing /l/ and /ɹ/ can be attributed to their heightened weighting of the F2 cue rather than the F3 cue.

Unlike Japanese listeners, Iverson et al. (2003) also demonstrated that German and English listeners maintain sensitivity to the F3 cue for the English /ɹ/-/l/ categorization and do not encounter significant difficulties in perceiving English /l/ and /ɹ/, as in both German and English, F3 is the primary cue to distinguish alveolar approximants /l/ and /ɹ/. Ingvalson et al. (2012) further investigated whether native Japanese listeners can learn to differentiate English /ɹ/ and /l/ based solely on F3 after training the participants for this differentiation. Thus, Ingvalson et al. (2012) further implied that, despite various efforts to teach this differentiation, native Japanese speakers typically do not achieve native-like performance. Together, these studies suggest that the listener's perceptual system can be tuned by the native language to prioritize cues that are unhelpful (e.g. F2 for English /ɹ/-/l/ categorization) for distinguishing between certain non-native sounds, thus impacting their ability to perceive these non-native phonemes correctly.

As noted earlier, these examples illustrate listeners' specific sensitivity to particular acoustic properties or cues of phonemes or for phoneme categorization, such as F0, F1, F2, and F3. Instead of these acoustic features, the present project utilizes this cue weighting mechanism to describe the weighting of cues that are conveyed at word- and sentence-level, such as the prosodic cues, verb semantic information and word information structure. However, the mechanism employed remains consistent. For the cues that are conveyed at word- and sentence-level, listeners exhibit preferences and sensitivity to certain cues during real-time speech processing. For instance, in the study by Scharenborg et al. (2020), non-native listeners exhibited a preference for distributed prosodic cues such as F0 over local cues like duration and intensity, leading to improved performance in non-native prosodic perception. French listeners, who rely more on F0 than on duration and intensity to perceive accent, demonstrate performance levels comparable to native English listeners when processing English sentences. Additionally, French and Finnish listeners, who

depend more on distributed prosodic cues in their native languages, outperform listeners who rely more on local cues like Dutch and English when processing non-native speech under noisy conditions. Consequently, this indicates that listeners who have a robust perception of a certain cue in their native language might transfer this cue weighting in L2 speech processing. This transfer could either facilitate or weaken the processing. The examples provided shed light on how listeners can leverage their native cue weighting of word- and sentence-level cues to non-native speech processing.

5.1.4 Other explanations: cognitive load and information integration

Many previous studies (Akker & Cutler, 2003; Hopp, 2010; McDonald, 2006) also mentioned that the cognitive load involved in real-time comprehension, compounded by the additional effort required to process and integrate non-native linguistic information, can overwhelm the available cognitive resources of the non-native listeners. The inability and difficulty for non-native listeners to utilize and integrate multiple cues is not necessarily due to a lack of knowledge but rather to the high cognitive and computational demands of processing language in real time. Thus, this strain limits their ability to access or integrate multiple cues effectively during on-line processing, affecting their comprehension and performance in non-native speech processing, and also affecting the ability to anticipate upcoming linguistic information.

For example, Akker and Cutler's study (2003) showed that non-native listeners tend to abandon prosodic processing during non-native speech perception. Akker and Cutler (2003) employed Dutch native listeners learning English as their L2 and listeners were presented with English sentences and were required to press a button as soon as they heard a word beginning with a specified phoneme. The target phoneme in each sentence was the initial phoneme in the target word, and the target word could be either accented or deaccented. For example, *The bones of the DINOSAUR were found by the Cuban archaeologist*, the word *dinosaur* was accented, and in the sentence *The bones of the dinosaur were found by the CUBAN archaeologist"*, the target word *dinosaur* was deaccented. The results showed that, unlike native listeners, non-native listeners did not demonstrate the same level of efficiency in using prosodic cues to anticipate the location

of accent in a sentence or react significantly faster to accented targets than to deaccented targets. Therefore, in Akker and Cutler's study, they revealed that the interaction between accent position and prosody facilitated English native listeners in swiftly locating the target based on sentence prosody cues. However, these effects did not manifest for Dutch native listeners when processing English sentences, indicating their inability to integrate sentence prosodic contours with discourse information.

Similarly, some studies (Ip & Cutler, 2020; Reed & Michaud, 2015) suggested that non-native listening may be challenged by an additional functional load, and non-native listeners generally lack awareness regarding prosody. These two features indicate that non-native listeners often struggle to integrate prosodic processing with other aspects of speech structure, and they also find it difficult to transfer skills from their native languages. For example, in Ip and Cutler (2020), Mandarin native listeners did not employ a prosodic entrainment strategy when processing English sentences. In the phoneme monitoring experiment of Ip and Cutler (2020), the difficulty in L2 listening may be exacerbated by the additional functional load during speech perception. In this study, Ip and Cutler employed native Mandarin listeners listening to English sentences. The English sentences were presented in two conditions: one with high stress on the target word carrying the target phoneme voiceless aspirated bilabial stop [p^h], such as the target word *party* in the sentence *I wish he weren't going to a PARTY on Monday*, and the second with low stress on the target word, such as *I wish he weren't going to a party on MONDAY*. The results indicated that when Mandarin listeners processed Mandarin sentences, the stress on the target word always facilitated the recognition of the target phoneme. However, when they processed English sentences, the stress on the target word did not facilitate the recognition of the target phoneme. Prosodic processing could be "abandoned" by Mandarin listeners when dealing with English sentences (Ip & Cutler, 2020). The presence of accent or stress patterns in English did not significantly aid native Mandarin listeners in accurately identifying the target phoneme. This suggests that the phonetic cues and prosodic features that native Mandarin listeners are accustomed to in their own language may not translate effectively to the perception of English speech.

Therefore, the studies discussed above show that non-native listeners, unlike their native counterparts, could struggle to process prosodic cues during real-time sentence comprehension. These studies underscored the cognitive and functional load that non-native listeners endure during real-time comprehension, significantly impacting their ability to process and assimilate non-native linguistic cues. This challenge isn't rooted in a lack of linguistic knowledge but stems from the intense cognitive and computational demands required for processing language in real-time. The challenge of employing and integrating additional cues, thus increasing cognitive and functional load, may partly explain the diminished proficiency in real-time sentence processing observed among non-native listeners.

Another challenge that non-native listeners face is their potential inability to fully benefit from both bottom-up input and top-down prior knowledge in their L2 at the same level as native speakers in speech perception (Ezzatian et al., 2010; Kaan, 2014). The utilization of the top-down information for native and non-native listeners during the low-level speech perception, has been discussed in the *Chapter 2*. During the low-level speech perception of perceiving speech units, the utilization of top-down information also emerges, raising questions about how these units are stored and whether phoneme-level units are stored in cognition. When it comes to high-level speech processing and the utilization and integration of various cues, the effective combination of both bottom-up and top-down information still remains essential.

Native listeners exhibit remarkable proficiency in skillfully utilizing both bottom-up cues, including factors like competing sounds and sound amplitudes, and top-down cues, such as prior knowledge about the speaker's voice, sounds, or words that they are going to listen to. Even under adverse backgrounds, native listeners' high familiarity of native sounds as bottom-up signals and efficient utilization of prior knowledge, optimize the efficiency of speech perception while concurrently minimizing potential disruptions or interference. This ability enables them to comprehend speech by extracting phonetic information from a target sentence and effectively discerning it from an adverse background, such as noisy environments where phonetic cues may be obscured by various sources of sound (Ezzatian

et al., 2010; Heald & Nusbaum, 2014; Scharenborg et al., 2016; Shuai & Gong, 2014; Sohoglu et al., 2012; Tabri et al., 2010).

On the other hand, information flow also plays an important role in forming predictive process during non-native language processing. Similar to native listeners, non-native listeners rely on top-down information to form a memory trace for the muted expected word (Foucart et al., 2016). Specifically, Foucart et al. (2016) used a lexical recognition task in which listeners first listened to a set of sentences where the target words were muted. The listeners then participated in a recognition task, where they were presented with a list of words—including expected (muted target word), unexpected (word that did not fit the context), old (words presented during the listening task), and new words (words not presented during the listening task)—and had to respond with "yes" or "no" to indicate whether each word had been present in the listening phase. Interestingly, expected words were recognized "yes" significantly more frequently than unexpected words, despite both being muted and this recognition was false as listeners should respond "no." Further, among the foils, expected words were more commonly misidentified as words listeners believed they had heard before, compared to unexpected words. Foucart et al. (2016) showed that non-native listeners demonstrated comparable performance to native listeners in utilizing top-down information to anticipate bottom-up information.

Foucart el at (2016) described this processing of using the top-down information to anticipate the word as: when bottom-up information is presented, listeners generate a "template" for the forthcoming information, whereas the absence of bottom-up information results in an empty "template," leading to a perceptual "hallucination" of the word. Consequently, listeners tend to recognize anticipated words as familiar more frequently than unexpected words. These findings affirm the existence of anticipation processes in non-native speech perception. Thus, it substantiates the notion that, similar to native speech perception, listeners engaging in non-native speech perception employ predictive processing and utilize top-down information to anticipate the upcoming target word, in which the top-down information could be low-level prosodic information or high-level pragmatic, morphosyntactic and semantic information. However, Foucart el at (2016) also noted that

non-native listeners may not exhibit the same degree of engagement as in native speech processing (Foucart et al., 2016).

In examining how listeners utilize and integrate cues, hypotheses and theories provide explanations for the diverse strategies employed in cue utilization and integration during native and non-native real-time sentence processing. Beyond cue integration, challenges such as phonetic distinctions, limited vocabulary, reduced lexical familiarity, and unfamiliarity with idiomatic expressions hinder non-native speakers' efficiency in understanding spoken language (Dijkgraaf et al., 2017). Previous studies typically focus on manipulating a few language cues in sentences to observe listener responses during real-time sentence processing; none have concurrently employed multiple cues to explore listeners' prioritization of particular cues. Although non-native listeners can use prosodic cues (Foltz, 2021; Perdomo & Kaan, 2021; Rosenberg et al., 2010; Tsurutani et al., 2010; Wagner, 2005) and verb semantics (Chambers & Cooke, 2009; Dijkgraaf et al., 2017; FitzPatrick & Indefrey, 2010) for prediction, the allocation of attention when these cues from both suprasegmental and discourse levels merge remains uncertain. This prompts an inquiry into the domain of cue utilization and integration, aiming to illuminate the overarching question in the field of speech processing: whether a preference for specific cues exists among native and non-native listeners; and whether this sensitivity to cues varies when processing new and repeated information structures.

5.2 Attention weight as an indicator of perceptual flexibility in sentence processing

Regarding the working principle of the Generalized Context Model (GCM), the concept of Attention weight plays a pivotal role within the GCM framework (Nosofsky, 1992; Nosofsky, 1986, 1988). It describes how attentional mechanisms enable listeners to focus on specific parts of the input sequence. Attention weight, as a parameter in speech models, was later applied by Johnson (1997) in speech perception to describe listeners' sensitivity to particular auditory properties when categorizing sounds, such as acoustic features or linguistic units.

However, in this thesis, rather than describing auditory properties, attention weight is

used to describe listeners' preferences and sensitivity to cues at the sentence level, such as semantics, prosody, and information structure, during their utilization and integration. Therefore, attention weight is a crucial concept and parameter in the eye-tracking experiment of this thesis.

According to Nosofsky and Johnson, attention weight describes the structure of psychological space during the process of stimulus categorization (K. Johnson, 1997; Nosofsky, 1992; Nosofsky, 1986). It can either contract or expand the perceptual space along each auditory dimension, aiming to enhance listeners' classification performance. Essentially, attention weight signifies the importance of auditory properties or dimensions in the categorization process, indicating which dimensions are more influential. Higher values of attention weight on specific dimensions lead to the expansion of the psychological space along those dimensions, suggesting their high relevance to the categorization process, with listeners being more sensitive to those dimensions during categorization. Conversely, lower values of attention weight cause the psychological space to shrink along particular dimensions, indicating their lack of relevance. This stretching and shrinking of psychological space directly impact the classification of new stimuli and influence the degree of similarity between the input stimulus and existing exemplars (K. Johnson, 1997; Nosofsky, 1992; Nosofsky, 1986, 2011).

In speech perception, Johnson (1997) provided an example of the attention weight assigned by listeners to multiple auditory properties during vowel categorization. According to the principle of attention weight, each auditory property represents a dimension, such as F0, F1, F2, duration, intensity, etc. For each perceptual segment, which can be viewed as a square, several auditory properties might exist, resembling grids on the square. Among multiple dimensions, F1 is the most crucial auditory property for distinguishing between vowels, as listeners categorize vowels primarily based on F1. Therefore, the F1 dimension will receive more attention weight when listeners categorize a set of vowel stimuli. Consequently, the perceptual space of this F1 dimension will expand, as listeners are most sensitive to this dimension and find it most relevant to the perception, while other auditory properties will contract as these dimensions are relatively less relevant than F1.

Therefore, in speech perception, the concept of attention weight initially examined how listeners allocate their attention, showcasing their perceptual flexibility, particularly towards sublexical units (e.g. acoustic properties, as summarized in Section 5.1.3 *Cue weighting theory*). However, as mentioned earlier, this thesis applies the concept at the sentence level. In this thesis, attention weight is a key indicator of perceptual flexibility in the utilization and sensitivity to particular cues during real-time sentence processing.

As mentioned in Chapter 1 *Introduction*, I posit that listeners with higher perceptual flexibility are adept at allocating attentional weights to various cues, demonstrating an enhanced capacity to utilize and integrate different cues adaptively to accommodate changes in the signal. Conversely, listeners with lower perceptual flexibility tend to exhibit rigidity in their attention allocation strategies. Such individuals often rely on a fixed approach, typically favoring certain cues over others, and may find it challenging to adapt by seamlessly integrating these cues in response to varying signal conditions. Reduced flexibility could underlie the observed differences between native and non-native sentence processing.

In the visual world eye-tracking experiment of this thesis, I examine the relationship between perceptual flexibility and attention weight: specifically, how the distribution of attentional weight across various cues reflects perceptual flexibility during real-time speech processing. This eye-tracking study uses attention weight as a metric to characterize how attention is directed toward specific cues within the auditory input. In the realm of sentence processing, as discussed in preceding sections, it is apparent that listeners engage with cues differently during native and non-native real-time speech processing: native listeners demonstrate a greater propensity to effortlessly and automatically utilize and integrate various cues (Cutler, 2012; K. Ito & Speer, 2008; Patterson et al., 2017; Pisoni, 1981), while non-native listeners exhibit a diminished preference for morphosyntactic and prosodic cues (Akker & Cutler, 2003; Clahsen & Felser, 2006a; Hopp, 2010; Jiang, 2007; McDonald, 2006), often encountering challenges in integrating cues from different domains (Patterson et al., 2017; Perdomo & Kaan, 2021).

Therefore, the current eye-tracking experiment aims to investigate the distribution of attentional weights across semantics, prosody, and information structure for listeners in native and non-native sentence processing. Specifically, this study examines word-level verb semantic cues, sentence-level information structure, pitch accent as prosodic cues, and the relationship between information structure and pitch accent. The investigation seeks to elucidate the processing dynamics of both English and Mandarin speakers in both their native and non-native languages.

5.3 Utilization and integration of semantic cues

5.3.1 Semantic predictive processing within visual-world eye-tracking paradigm

Undoubtedly, predictive processing of upcoming elements plays a crucial role in real-time speech processing and comprehension, requiring listeners to effectively utilize word- and sentence-level cues, such as semantic and syntactic information, for effective prediction. Listeners leverage semantic cues, particularly from verbs, to anticipate forthcoming elements during speech processing, underscoring the significance of verb semantics in the predictive process. On the one hand, the verb semantic cue also plays a significant role in enabling listeners to anticipate upcoming nouns or objects during real-time speech processing. On the other hand, understanding the nature of the predictive process illuminates how cues are utilized and integrated by listeners in both native and non-native contexts. According to Kaan (2014), predictive processing in non-native sentence processing is proposed to not inherently diverge from predictive processing in native sentence processing. However, it may be influenced by variables linked to non-native understanding. Thus, essentially, employing predictive processing serves as a gauge or means to reflect the listeners' ability to utilize and integrate cues effectively. There exists a close and profound connection between the utilization of verb semantics and predictive processing in real-time sentence processing.

This project specifically focuses on verb semantic information as a critical cue to assess attention allocation among listeners in both native and non-native real-time speech perception contexts. In this chapter, I review previous studies that elucidate the role of

semantic cues, with a special emphasis on the semantic cues at verbs, and their impact on the prediction and anticipation processes during real-time speech perception. This analysis is conducted within the framework of the visual-world eye-tracking paradigm, highlighting how these cues enhance the understanding of native and non-native listener engagement and cognitive processing in real-time spoken language.

The concept of anticipation in speech perception refers to the process of predicting what comes next in a sentence based on the contextual cues and previous linguistic and non-linguistic information. Anticipation during real-time speech perception manifests as the connection between linguistic elements in sentences and real-world events. Therefore, understanding the essence of anticipation during speech perception is crucial for comprehending how verb semantic cues facilitate anticipation during sentence processing for both native and non-native listeners. As reviewed in Altmann and Mirković's study (2009), the model *simple recurrent network* (SRN), originally proposed by Elman (1990) and later developed by Dienes et al. (1999), describes the mechanisms of anticipation in speech perception. This SRN model posits that the process of anticipating upcoming elements of language plays a pivotal role in understanding and processing both linguistic and non-linguistic information. Therefore, based on the working principles of this model, Altmann and Mirković (2009) outlined four key principles to characterize the language comprehension: (1) linking across different domains; (2) prediction; (3) contextual influence; and (4) processing linguistic structure over time spans.

According to the illustration of Altmann and Mirković's work (2009), firstly, language comprehension during speech perception is an active cognitive process of bridging various domains, especially bridging between the sentence structures and the external world structures, such as the real-world events. The elements of sentence structures such as semantic and syntactic elements is crucial in facilitating listeners to predict and understand the structure of events in the external world. Therefore, language comprehension is not a passive task for listeners to simply recognize words and morphosyntactic structure. Instead, it is an active process for listeners to establish a cognitive connection between the evolving language and the real-world events it represents. In other words, comprehension

occurs when listeners establish connections between linguistic structures and real-world events. Secondly, listeners utilize their linguistic knowledge to anticipate forthcoming information during speech perception. Thus, the predictive process during the speech perception plays a crucial role in understanding and incorporating the connection between language and the external world. Thirdly, the predictive process during the speech perception is shaped, influenced and driven by the contextual cues. The context is not only composed of the linguistic information such as the segmental and suprasegmental information, but also composed of the nonlinguistic information, such as the visual information, physical surroundings, and background sounds, etc, and the prior internal states of listeners, such as personal experience and memory, attention, working memory, emotional states, etc. Fourthly, the predictive process extends across different time frames and different levels of conceptual abstraction. Namely, the predictive process can extent from predictions in the short term to the context in the long term and from specific details to broader ideas. Therefore, the utilization of semantic cue within predictive processes is crucial in listeners' comprehension of the mapping between language structure and real-world events.

To manifest the mapping between the sentence structure and the external-world events, the visual eye-tracking paradigm is an important approach for demonstrating the anticipatory eye movements during language processing. Eye movements are influenced by the dynamic interaction between the language and the mental representation of the scene. As a result, the eye movements of individuals are not solely directed by visual input, but are also influenced by a continuously updated mental representation of the external world based on the received new information and the changes of individual's understanding of the external world. According to Altmann and Mirković's review (2009) on the significant role played by the visual-world eye-tracking paradigm in visualizing the language comprehension and processing, the prediction manifested by the visual-world eye-tracking paradigm falls into the following three domains: first, the eye-tracking paradigm manifests the predictive process of upcoming information within the linguistic domain, such as predicting the upcoming phonological and orthographic information; second, the predictive process can extend into the visual domain, such as predicting the upcoming visual

objects that will be mentioned next; third, the predictive process can also operate in the conceptual domain, such as predicting the upcoming relevant mental representations.

In the context of predictive processes within the linguistic domain, such as in the context of anticipating the phonological or orthographic forms of upcoming nouns, particularly in native real-time speech processing, listeners demonstrated an ability to anticipate the phonological and orthographic information of forthcoming nouns based on preceding semantic context (A. Ito & Sakai, 2021; A. Ito et al., 2018). However, unlike native listeners, non-native listeners do not exhibit specific anticipation of the phonological information associated with the upcoming noun.

For example, Ito et al. (2018) conducted a visual-world eye-tracking experiment, involving both native and non-native listeners: English native listeners and Japanese native listeners learning English as L2 (Japanese-English listeners). The study presented English sentences like *The tourists expected rain when the sun went behind the ...*, and asked listeners to observe an array of objects (e.g. cloud, clown, bear, globe). These objects include a target object whose English name corresponded to the predictable word (e.g. *cloud*, with the corresponding Japanese translation "kumo"), an English competitor object whose English name was phonologically related to the predictable word (e.g. *clown* as English competitor object phonologically related to the target object *cloud*, with the corresponding Japanese translation being "piero") , a Japanese competitor object whose Japanese name was phonologically related to the Japanese translation of the predictable word (e.g. *bear* as Japanese competitor object, its Japanese translation is "kuma" which is phonologically related to the Japanese translation "kumo" for the target object *cloud*), and an object that was unrelated to the predictable word (e.g. *globe* as unrelated object, its Japanese translation is "tikyuugi"). As the stimuli were in English, the results showed that both native and non-native listeners exhibited predictive looks towards the target object, although Japanese-English listeners demonstrated a delay in predictive eye movement on the target object compared to native listeners. However, only native English listeners demonstrated the ability to predict the phonological information of the upcoming noun, as evidenced by their predictive looks at the English competitor object (e.g. "clown") whose

English name was phonologically related to the target object (e.g. "cloud"). Compared to native listeners, Japanese-English L2 listeners did not show the predictive looks to the English competitor object (e.g. "clown"), suggesting that in non-native real-time sentence processing, listeners were not able to predict the specific phonological information of the upcoming noun. Moreover, Japanese-English L2 listeners neither showed the predictive looks towards the Japanese competitor object (e.g. "bear") nor exhibited a preference for the Japanese competitor object over the unrelated object, indicating that L1 phonological information (e.g. Japanese) would not be activated during L2 (e.g. English) comprehension and processing. Thus, listeners can predict phonological information in their native language but may not do so in a non-native language.

Another study by Ito and Sakai (2021) used Japanese stimuli and Japanese native listeners to test native listeners' ability in predicting the orthographic forms of a highly predictable word during listening comprehension. They used the critical object represented the target word *(e.g. sakana "fish")*, an orthographic competitor *(e.g. tuno "horn")* which is written in a similar form as *sakana "fish"* in Japanese, a phonological competitor *(e.g. sakura "cherry blossom")*, or an unrelated word *(e.g. hon "book")*. Listeners listened to Japanese sentences like *"This is a fish that my father caught in the ocean nearby, a type of which is rarely found in supermarkets"*. Because the stimuli were in Japanese, whose word order differs from English, the target object "fish" appeared near the end of the sentence. Listeners were then instructed to click on the object that was mentioned in the sentence. The results showed that before the target word was mentioned, listeners fixated the target and the orthographic competitor over the unrelated objects, suggesting that listeners activated the orthographic form of the target word in advance.

These studies underscore prediction as a widespread and fundamental aspect of real-time speech processing and comprehension, affecting multiple domains of representation, such as phonological and orthographic information. Nevertheless, due to disparities in performance between native and non-native listeners during real-time speech perception, predictive capabilities in specific domains of representation, like phonological or orthographic forms, may be exclusive to native listeners, as opposed to non-native listeners.

In the context of predictive processes within the visual and conceptual domain, this typically involves leveraging the information conveyed by verbs, combined with the listener's knowledge of the world, to anticipate forthcoming nouns or objects. The anticipation in this domain is related to conceptual domain knowledge and how listeners mentally represent the relationship between verbs and their possible noun complements. In contrast to the scenario of predicting phonological and orthographic forms, non-native real-time speech processing reveals listeners' ability to anticipate forthcoming nouns or objects by extracting semantic information from preceding verbs. However, previous studies indicated that non-native listeners exhibit a delay in providing predictive gazes towards the target object compared to native listeners.

The study of Altmann and Kamide (1999) employed English native listeners and English stimuli. The study used the visual objects to showed that how verbs and their subjects can be used to anticipate upcoming arguments. This eye-tracking experiment demonstrated that listeners possessed the ability to utilize verb information to anticipate the characteristics of forthcoming objects. The study used sentences like *"The boy will move the cake"* and *"The boy will eat the cake"*. Results showed that the listeners looked to the target object *(e.g. "the cake")* significantly slower when the verb was *"move"* compared to when the verb was *"eat"*. When the verb was *"move"*, the eye movements towards the target commenced after the word *cake* had been spoken, whereas when the verb was *"eat"*, these eye movements were initiated prior to the onset of the word *cake*. It suggests that the information conveyed by the verb in a sentence can affect and restrict how objects are later mentioned within the context of that sentence. In the second experiment in Altmann and Kamide's work, they used the same stimuli and procedure, but the listeners were instructed to ignore the auditory stimulus and simply focus on the visual stimuli. Therefore, the first experiment involves active sentence processing, while the second experiment focuses on visual attention when listeners were not actively engaged in sentence processing. They obtained the similar results, suggesting that information extracted from the verb can direct eye movements, irrespective of whether it involves active sentence processing. Consequently, this study provides support for the idea that in native real-time speech processing, listeners can leverage the predictive relations among verbs, their syn-

tactic structures, and the real-world contexts in which they are employed to comprehend sentences.

Other visual-world eye-tracking studies by Corps et al. (2023) and Chun and Kaan (2019) have showed that listeners can predict upcoming objects semantically related to the preceding verb, even in non-native real-time speech processing. For instance, in the study conducted by Corps et al. (2023), L2 English listeners were presented with sentences such as *I would like to wear the nice...* alongside four possible upcoming objects (e.g. tie, drill, dress, hairdryer) showed on the screen, where the *tie* and *dress* were target objects indicating masculine and feminine options, respectively, and semantically linked to the verb *wear*, while the objects *drill* and *hairdryer* were the distractors. The results indicated that English L2 listeners successfully anticipated both masculine and feminine target words by leveraging the semantic context provided by the preceding verb. Moreover, this study also incorporated the apparent gender of the speaker. Listeners listened to sentences produced by both male and female speakers and were more likely to fixate on objects that were stereotypically associated with the gender of the speaker after the verb was heard. This indicates that both native and non-native listeners were integrating speaker gender with semantic information from the verbs to predict the outcomes, reflecting an interaction between semantic processing and real-world knowledge. Similarly, Chun and Kaan (2019) conducted a study with English L2 listeners who listened to sentences like *"I know the friend of the dancer that will open/get the..."*, where the verb *"open"* served as a semantically-biasing verb, and the verb *"get"* was semantically neutral. The findings revealed that, akin to L1 listeners, English L2 listeners exhibited significantly more anticipatory gazes towards the target object *"present"* in the semantically biasing condition compared to the neutral condition. Although the predictive gazes of English L2 listeners commenced later than those of native English speakers.

In essence, the predictive processing observed in the visual-world eye-tracking paradigm is crucial mechanism for listeners' understanding of how language structures relate to real-world situations. Within this process, listeners not only predict the next linguistic elements in a sentence but also anticipate the real-world events to which those elements

refer to. Consequently, investigating the capacity of both native and non-native listeners to leverage semantic cues, particularly the semantic information at verbs, is essential in understanding predictive processing during real-time sentence processing. In other words, predictive processing within the visual-world eye-tracking paradigm serves as a pathway illuminating native and non-native listeners' ability and proficiency in utilizing cues.

Thus, the discussion focuses on the verb semantic information in the following sections is grounded in the alignment of the current project's eye-tracking experiment with the relevance of verb semantics which is used to investigate listeners' gaze patterns directed towards subsequent objects. Given the close association between the utilization of verb semantic cues and predictive processes during speech perception, the experiment in this project seeks to understand how listeners allocate attention weight across various cues. This understanding is informed by the interplay between predictive processing and the utilization of verb semantic cues, along with other cues.

5.3.2 Native listener's ability in using verb semantic cue

As mentioned in the preceding section, anticipation plays a central role in current approaches to sentence comprehension. Listeners do not simply passively integrate information to reconstruct linguistic meaning; rather, they actively construct interpretations by generating anticipations based on the integrated information (Pickering & Garrod, 2013). In addition to semantics associated with verbs, native listeners proficiently utilize semantic cues from pronouns and adjectives (Chambers et al., 2002; Sedivy et al., 1999) to predict upcoming nouns. They seamlessly integrate these cues with other speech elements, such as syntactic information and semantic information at other sentence elements (Kamide, Altmann, & Haywood, 2003; Kamide, Scheepers, & Altmann, 2003). For example, the semantics of adjectives also significantly contribute to enhancing the predictive abilities of native listeners. This phenomenon has been consistently demonstrated in various studies employing both the visual-world eye-tracking paradigm and ERP studies (Sedivy et al., 1999; Szewczyk et al., 2022).

This project primarily aims to investigate the significance of verb semantic information

as a cue at the sentence level. It is well-known that the verb semantics are commonly used by the sentence processor to make the prediction. This section focuses on reviewing previous studies that shed light on how native listeners utilize verb semantic information to predict forthcoming nouns. The utilization of verb semantics encompasses two critical components: firstly, the ability of native listeners to employ verb semantics to anticipate the characteristics, location, and state of the forthcoming noun. Secondly, native listeners demonstrate their proficiency in integrating verb semantics with other linguistic cues, such as semantic information at nouns or subjects, tense, and contextual information. This integration serves to refine and restrict their predictions concerning the upcoming objects.

In the following studies, listeners extracted the information at verbs while listening to sentences to predict the location of the upcoming objects in visual-world eye-tracking experiments. The expectation regarding the location of the forthcoming object is predominantly influenced by the information conveyed by the verb and the contextual information, rather than the actual object location in the scene presented to them. The study of Altmann and Kamide (2009) reported a visual-world eye-tracking experiment, in which with the visual scene unchanged, listeners listened to sentences either in a "moved" condition *"The woman will put the glass onto the table. Then, she will pick up the bottle, and pour the wine carefully into the glass"*, or in an "unmoved" condition *"The woman is too lazy to put the glass onto the table. Instead, she will pick up the bottle, and pour the wine carefully into the glass"*. In the visual scene presented to the listeners, several objects were featured: a woman, a bookshelf, an empty wine glass, and a table. Notably, these objects were arranged separately; in other words, the wine glass was not positioned on the table. Listeners were asked to look at the pictures while listening to sentences that described something related to the content of the pictures. Listeners were expected to make a predictive look to the table rather than the glass after hearing the verb *"pour"*, as the object table served as the location of glass of the pouring event. The experiment was designed to investigate how participants' eye movements were influenced by by verb semantic information and also in combination with the contextual information.

The verb *"pour"* suggests an action comprising a source, an activity, and a destination, such as the act of transferring liquid from a bottle into a glass in this context. The study revealed two things, first, in both the "moved" and "unmoved" conditions, listeners tended to look more at the table rather than the glass. This is because the verb *"pour"* and the contextual information provided in the sentences led them to anticipate the pouring action occurring at the table, despite in the "unmoved" condition the glass not being physically positioned there in the visual scene. Following the verb *"pour"*, the eye movements anticipated the goal of the pouring, which is where the wine would be poured, during the time window after the verb and prior to the noun *"the glass"*. This is consistent with the previous research (Kamide, Altmann, & Haywood, 2003) that the anticipatory eye movements tend to be directed toward the most plausible goal object following a ditransitive verb, in this context, which was the table. This indicates that listeners' eye movements were influenced by the information conveyed by the verb even when it contradicted the actual location of the glass, as the glass was not placed above the table in the presented scene. Second, starting from the word *"pour"* and continuing, more saccadic eye movements were directed towards the table in the "moved" condition compared to the "unmoved" condition. This suggests that listeners' eye movements were influenced not only by verb semantic information but also by the combination of contextual information provided in the sentences. In accordance with the contextual information in the sentence, in the "moved" condition, the glass was moved from the floor to the table. Consequently, listeners directed more eye movements toward the table more compared to the "unmoved" condition.

In the second experiment in the study of Altmann and Kamide (2009), a "blank screen paradigm" was employed. In this experiment, listeners were initially presented with a scene for a few seconds, after which the screen went blank. Following a brief pause of one second, the audio stimuli were played, accompanied by the reappearance of the scene. Consequently, listeners in this "blank screen paradigm" were aware of the glass's location (e.g. glass on the floor) in the scene before the sentence was presented. Nonetheless, even with prior knowledge of the glass's exact position, their gaze remained fixed solely on the mental location of the glass (the glass's location given in the context of the sentences

they heard, such as, *The woman will put the glass on the table...*). This fixation was determined by the interplay between the language and their prior knowledge. The actual physical location of the glass or other objects in the scene did not significantly impact the eye movements of the listeners. Therefore, it was evident that listeners successfully extracted information from the verb and context to predict the location of the upcoming noun, and their eye movements were primarily influenced by the language itself rather than the visual scene.

In addition to anticipating the upcoming object's location, native listeners were also able to utilize the verb information to straightforwardly predict what is the forthcoming objects (G. T. Altmann & Kamide, 1999; Arai & Keller, 2013; Lew-Williams & Fernald, 2007; Van Bergen & Flecken, 2017). For example, in the study of Lew-Williams and Fernald (2007), one of the experiments in this study examined online processing tasks involving verb thematic information. Native Spanish listeners were employed, and Spanish stimuli were utilized. The researchers used the "looking-while-listening" procedure, in which participants observed two pictures while simultaneously listening to speech that named one of the pictures. The eye movements of the participants were monitored to assess their ability to predict the upcoming noun following the verb. Listeners were exposed to sentences with contextually constraining verbs. Some sentences contained verbs that were semantically related to the following noun *Cómete la galleta; "Eat the cookie")*, while others had verbs that were unrelated to the following noun *(e.g Encuentra la galleta; "Find the cookie")*. As the listeners listened to these sentences, they viewed two corresponding images (e.g. a ball and a cookie). The selection of sentence and the timing of the appearance of the images were designed to assess the speed of listeners recognized the subsequent nouns in the sentences based on the verb information. Spanish listeners looked more toward the object "cookie" with a semantically related verb *eat* than with a semantically unrelated verb *find*. Native speakers of Spanish showed that they were able to use verb information for predictive processing during the comprehension of spoken language.

Similarly, the study conducted by Bergen and Flecken (2017) also investigated the role

of verb semantics in predicting upcoming objects in Dutch sentence processing. Native Dutch listeners participated in the experiment, during which they listened to Dutch sentences while simultaneously viewing visual displays featuring objects positioned in different locations. The participants' eye movements were recorded. The results of the study revealed that native Dutch listeners exhibited anticipatory eye movements towards objects that corresponded to the position implied by the placement verb. When listeners heard the verb *zetten ("to put.stand")*, they increasingly focused on standing objects. Conversely, when listeners heard the verb *leggen ("to put.lie")*, their proportion of fixations on standing objects decreased. These findings also strongly indicated that native Dutch listeners were proficient in extracting the positional information conveyed by verbs to predict upcoming objects, as evidenced by their anticipatory eye movements.

Moreover, it is interesting to note that native listeners were not only able to make a predicative look to the upcoming noun, but also able to anticipate the state of the upcoming nouns (G. T. M. Altmann & Mirković, 2009; G. T. Altmann & Kamide, 2007; Knoeferle & Crocker, 2007). In the study of Altmann and Kamide (2007), native English listeners were exposed to English stimuli. Listeners listened to the sentences like *"The man will drink the beer"* or *"The man has drunk the wine"* in the experiment 1, while viewing a scene featuring a man, a full glass of beer, an empty wine glass, some cheese, and some Christmas crackers. The results revealed that more saccades were directed towards the empty wine glass in the past tense conditions compared to the future tense conditions. Conversely, in the future tense conditions, more saccades were observed towards the full glass of beer. To get rid of the potential influence of whether the man might have drunk some of the beer, Experiment 2 maintained the same scene. However, listeners listened to sentences like *"The man will drink all of the beer"* or ''The man has drunk all of the wine", yielding similar results to Experiment 1. Likewise, the same study by Altmann and Kamide (2007) included another set of sentence pairs *"The cat will kill all of the mice"* and *"The cat has killed all of the birds"*. The scene included a cat, a few mice, and some feathers. When listeners listened to the sentence *"The cat will kill all of the mice"*, their gaze was primarily directed towards the mice rather than the feathers before the phrase *"the mice"*. Conversely, when listeners listened to the sentence *"The cat has killed all of*

the birds", their attention shifted more towards the pile of feathers than the huddled mice prior to hearing *"the birds"*.

In these two sentence pairs from Altmann and Kamide's study (2007), when listeners heard sentences like *"The man has drunk all of the wine"* and *"The cat has killed all of the birds,"* they directed their attention more towards the empty wine glass and the feathers compared to the other objects in the scene. It is worth noting that both the empty wine glass and the feathers violated the selectional restrictions associated with the verbs used, as pointed out by Altmann and Mirković. According to these restrictions, the upcoming nouns for the verbs *"drink"* and *"kill"* should be something drinkable and animate which can be killed by a cat. However, native listeners demonstrated an ability to associate the present perfect tense with the resultant states, indicating that after a bird has been killed by a cat, there should only be some feathers left on the ground. Altmann and Kamide (2007) also comments on their earlier study (1999) in which they utilized sentence stimuli in the future tense, such as *"The boy will eat the cake"*. In this scenario, listeners directed their attention towards the cake. However, if the present perfect tense *"has eaten"* were used, listeners may shift their gaze towards the empty plate. Therefore, these experiments suggested that native listeners are able to utilize semantics of tense to predict the state of upcoming nouns, and this prediction is based on the combination of both conceptual structures activated by language and the visual scene. Anticipatory eye movements of native listeners are driven by this overlap, reflecting an unfolding mental world rather than simply the unfolding language itself.

Another noteworthy point is that native listeners demonstrated the capability to effectively utilize and integrate semantic information from both verbs and nouns for predictive purposes. The study of Hopp (2015), which has been reviewed in the previous section (*Section 5.1*), German native listeners were found to integrate morphosyntactic and lexical-semantic information in anticipatory processing. In the study of Kamide et al. (2003), the first experiment also illustrated how native listeners proficiently incorporate semantic information from verbs to make predictions. Native English listeners were engaged, and English stimuli were employed. Listeners were instructed to listen to sentences like

"The woman will spread the butter on the bread" and *"The woman will slide the butter to the man"* while viewing images on the screen. The results of this study indicated that native English listeners exhibited predictive eye movements towards the bread, which is the goal in this context. Likewise, listeners displayed predictive eye movements towards the man when the verb *"slide"* was heard. In their second experiment, this study revealed that prediction is not only influenced by the verb but is also constrained by the subject preceding the verb. Native listeners exhibited the ability to rapidly integrate information from both nouns and verbs to form predictions. The study sought to confirm whether the information about the subject, referred as the *Agent* in Kamide et al's study, could combine with the selectional restrictions of the verb to predict the direct object, referred as the *Theme* in their study. They used the sentences like *"The man will ride the motorbike"* and *"The girl will ride the carousel"*. The results of experiment 2 showed an increased eye movements to potential direct objects of the verb *(e.g. "the motorbike")* after listeners heard sentences like *"The man will ride"*. This is based on the plausibility from the real world. It's a reasonable expectation that a man would be inclined to ride a motorbike, while a young girl would typically choose a carousel. Additionally, the results suggest that the increased looks to the object cannot be attributed solely to the verb's selectional constraints. Therefore, native listeners demonstrate the ability to anticipate the semantic characteristics of upcoming direct objects, and this is based on the combination of subject information and the verb information, providing insights into the rapid integration of the information at both nouns and verbs during sentence processing.

5.3.3 Non-native listener's ability in using verb semantic cue

Like their native counterparts, non-native listeners are capable of extracting semantic meanings from individual words to predict forthcoming target nouns. Non-native Mandarin listeners can deduce semantic meanings from measure words (modifiers) to predict subsequent nouns (Theres Grüter & Ling, 2020). Non-native listeners are similarly skilled in using preceding verb semantics to forecast the following noun.

Chambers and Cooke's eye-tracking study (2009) involved English native listeners learning French as L2 and demonstrated that listeners could extract semantic meaning from

verbs to eliminate cross-language lexical competition and constrain the subsequent target noun. The study used French stimulus sentences such as *Marie va décrire la poule* ("Marie will describe the chicken") and *Marie va nourrir la poule* ("Marie will feed the chicken"). The four objects presented on the screen included a target noun (e.g. a chicken), a cross-language competitor (e.g. a pool), and two distractors (e.g. a strawberry and a boot). The results revealed that, when presented with French target noun *poule* ("chicken") in the sentence such as *Marie va décrire la poule* ("Marie will describe the chicken"), where the verb *décrire* ("describe") was semantically fitting for both the target word *poule* ("chicken") and the English competitor *pool*, English native listeners momentarily considered competitor English words, such as *pool*, which is a near-homophone of the French target nouns, such as *poule* ("chicken"). In contrast, when presented with French target noun *poule* ("chicken") in the sentence such as *Marie va nourrir la poule* ("Marie will feed the chicken"), where the verb *nourrir* ("feed") strongly constrained the upcoming target word to be *poule* ("chicken") and was semantically incompatible with the English near-homophone competitor word *"pool"*, the impact of cross-language lexical competition was notably diminished. Consequently, English native listeners showed a dramatic decrease in fixations on the interlingual English competitor words.

This affirms that, during non-native speech perception, listeners can efficiently extract semantic cues from the verb to predict the upcoming noun and mitigate cross-language lexical competition from their native language. This aligns with Ito et al.'s study (2018), summarized in *Section 5.3.1 Semantic predictive processing within visual-world eye-tracking paradigm*. Recall in that study, Japanese native speakers learning English as L2 did not show predictive looks towards the Japanese competitor object when listening to English sentences, indicating that listeners did not activate their L1 Japanese phonological information during L2 sentence processing. Non-native listeners demonstrated the ability to mitigate cross-language lexical competition from L1.

Similarly, Dijkgraaf et al.'s eye-tracking study (2017) involved Dutch native listeners who were learning English as an L2, as they listened to English sentences. The stimulus sentences, such as *"Mary reads a letter"* (constraining condition) and *"Mary steals a*

letter" (neutral condition), served as examples. Eye-movements were recorded as listeners viewed a screen displaying four objects, such as a letter, a backpack, a car, and a wheelchair. All these four objects could be "stolen" (neutral condition), but in which only one object (the letter) could be "read" (constraining condition). Dijkgraaf et al. predicted that if listeners could predict the upcoming target noun based on the verb's semantics, there would be a higher proportion of fixations on the target object (e.g. the letter) upon hearing a sentence like *"Mary reads a letter"* compared to hearing a sentence like *"Mary steals a letter"* before the target was audibly presented. They also hypothesized that non-native listeners would show a delayed and reduced focus on the target object relative to their monolingual counterparts. The findings indicated that English L2 listeners, during non-native speech perception, exhibited a predictive process akin to native speakers, albeit with a delayed focus. This confirms that listeners can extract semantic information from the verb to anticipate the subsequent noun to a degree similar to that observed in native speech processing.

Typically, in addition to relying on specific words such as verbs to predict the upcoming noun, it's important to recognize that the overall semantic context of a sentence plays a crucial role in real-time speech processing. This context significantly influences listeners, including non-native ones, guiding them to effectively utilize the broader semantic framework of the sentence to refine their anticipation of the target word. Thus, it is also noteworthy that non-native listeners have demonstrated the ability to utilize the overall semantic context of the sentence to constrain the anticipated word (FitzPatrick & Indefrey, 2010; Lagrou et al., 2013). Previous ERP studies consistently indicate that non-native listeners can utilize the sentence context to enhance word prediction (Chambers & Cooke, 2009; FitzPatrick & Indefrey, 2010; Foucart et al., 2016; Lagrou et al., 2013): specifically, when a word is incongruent with the sentence semantics, non-native listeners exhibit a significant N400 effect. Conversely, when the word aligns with the sentence semantics, a diminished N400 effect is observed in non-native listeners.

As mentioned previously, non-native listeners exhibit delays in predictive processing compared to native counterparts. Additionally, they face challenges in integrating seman-

tic cues with other cues to make predictions. Non-native listeners encounter difficulties in integrating morphosyntactic and lexical-semantic information during the anticipation process, although they demonstrate an ability to extract lexical-semantic details solely from verbs to predict the semantics of upcoming input (Hopp, 2015) (discussed in Section 5.1). Further, as stated in the eye-tracking study of Mitsugi and Macwhinney (2016), while listeners have showed the ability to utilize semantic cues during real-time speech perception, indicating that non-native listeners may possess strong offline knowledge of their non-native language, but they could encounter difficulties in effectively applying this knowledge in real-time during visual-world eye-tracking experiments, inhibiting their ability to generate predictions swiftly. To be specific, Japanese applies a Subject-Object-Verb (SOV) structure with the case marking system to indicate sentence components. This special case marking system in Japanese is considered to impose additional processing costs for non-native listeners (Nakano et al., 2002). Mitsugi and Macwhinney (2016) employed non-native Japanese listeners and first asked to do a grammar task to make sure these non-native listeners of Japanese understand the the Japanese case markers *"-ni"* (indicating the indirect object of a verb or the destination of a movement), *-ga* (indicating the subject or nominative case), *-o* (indicating the direct object) and *-de* (indicating the location).

The example sentence given in Mitsugi and Macwhinney's study was like:

(3) *gakkou-de majimena gakusei-ga kibishii sensei-ni shizukani tesuto-o watashita.*
school-LOC serious student-NOM strict teacher-DAT quietly exam-ACC handed-over.
'At the school, the serious student quietly handed over the exam to the strict teacher.'

Listeners were asked to listen to sentences and looked at visual scenes containing direct and indirect objects (e.g. both "exam" and "teacher") that the sentences described. The results suggested that, despite showing good offline knowledge of Japanese case markers, during real-time non-native speech perception, non-native Japanese listeners seem unable

to apply, or at least not quickly enough, some knowledge in their L2 to facilitate prediction. Therefore, difficulties arise in the real-time application of offline knowledge in this context. Non-native listeners of Japanese did not exhibit the same anticipatory eye movement to the direct (e.g. "exam") and indirect objects (e.g. "teacher") as native speakers, indicating a significant challenge in employing case markers for predictive processing during listening tasks.

5.4 Utilization and integration of pitch accent in relation with information structure

It is widely recognized that cross-linguistically, pitch accent serves as a crucial suprasegmental or prosodic cue utilized in speech to emphasize or highlight specific words or syllables (Cole, 2015; P. Warren, 1999; Wenrich et al., 2017). Because of pitch accent's inherently combinatorial purpose (e.g. pitch accent interacts with information status, sentence and word meanings), prosody offers the perfect type of cue to examine when looking at how listeners integrate cues. In both English and Mandarin, the fundamental frequency (F0) plays a critical role in conveying focus by influencing the perceived pitch. Typically, syllables that carry pitch accents exhibit a higher F0 or more pronounced fluctuations in F0 compared to those without accents (Roessig et al., 2022). Typically, a higher or greater F0 is associated with increased and heightened prominence.

Building on the understanding that pitch accent plays a pivotal role in highlighting specific words or syllables, it closely intersects with the concept of *information structure* in speech: the combinatorial purpose of pitch accent sheds light on the information structure of discourse. The concept of information structure was originally introduced and coined by Halliday (1967). It has since been closely associated with differentiating given and new information expressed in sentences (Gundel, 2012). According to Halliday's (1967) concept of information structure, it pertains to how languages encode and organize the informational status of sentence components. It involves how sentences are built to enhance the efficient processing of information based on the requirements of participants in the communication.

Therefore, information structure plays a pivotal role in effectively conveying meaning and ensuring the efficiency and effectiveness of communication. The terms *information status* and *information structure* are often used interchangeably under many circumstances. A sentence's information structure segregates into given and new information. Comprehending the sentence involves linking the given information with what is already known and incorporating the new information into the ongoing mental model of the discourse In the simple case we consider here, a sentence is divided in its information structure into given and new information. Understanding the sentence requires connecting the given information to information already available and adding the new information into the evolving mental representation of the discourse (Clark & Haviland, 1977).

However, it's important to note that information structure primarily pertains to the concept of *relational givenness-newness* (Topic-Focus structure) (Gundel, 2012; Gundel & Fretheim, 2006), which is a broader concept that forms the information structure. It deals with the relationship between the topic (what is given in the discourse) and the focus (what is new or asserted). In comparison, referential givenness-newness is a more specific concept derived from the relational framework, describing whether entities in the communication are already given or new in the discourse. For example:

(1) A new student arrived. She seemed nervous.

(2) Yesterday (Topic), I met my old friend (Focus).

Sentence (1) refers to referential givenness-newness, the *new student* represents referentially new information, while *she* is referentially old information in this context. On the other hand, sentence (2) refers to relational givenness-newness, *yesterday* serves as the topic, establishing the temporal context as old information, while *I met my old friend* function as the focus, introducing new information about the event.

According to Gundel and Fretheim's explanation in their work (2006), the relational givenness-newness describes the thematic structure of sentences, centering around the concepts of *topic* and *focus*. The topic and focus function as a pair of complementary elements, noted as X and Y. X represents the subject matter of the sentence, mainly concerning with what the sentence is about; while Y is the predication of X, conveying information related to X. In terms of this, X represents the given information in relation to Y, and Y introduces new information in relation to X. In sum, *information structure* articulating and differentiating between new and previously established information within sentences, draws its origins from the broader concept of relational givenness-newness.

5.4.1 Pitch accent as prosodic cue

5.4.1.1 Pitch accent in relation with information structure

Pitch accent is considered as prosodic information. One of its functions is to convey information structure, which is considered as sentence-level information during the real-time speech processing. According to prosodic theory (Wells, 2006), the positioning of the sentence nucleus is markedly influenced by whether the words convey previously established or new information. Typically, accentuation indicates new information, while deaccentuation signifies old or reiterated information. In this thesis, the information structure typically refers to the referential givenness-newness. As mentioned previously, the combinatorial nature of pitch accent closely interacts with the information structure. The expected pitch accent can vary based on whether a target word represents previously given or new information. In other words, the prosodic structure plays a vital role in conveying information structure at the discourse level (Ouyang & Kaiser, 2015; Watson, 2010). In this thesis, *information structure* and *contrastive pitch accent* are employed as both sentence-level information and prosodic information to examine how listeners allocate their attention during real-time speech perception.

One of the uses of pitch accent as a prosodic cue is to convey information structure. New information is typically accentuated as *focus* (prominence), while old information is usually not emphasized (accentuated) (Halliday, 1967). The focus is achieved by employing pitch accents to highlight specific elements within sentences, referring to the prominence

in the intonational context (Birch & Clifton, 1995). Listeners quickly and accurately discern that a subsequent sentence in a dialogue is logical when the new information is highlighted and the given information is not (Birch & Clifton, 1995). Early research has demonstrated that accents signal that a new discourse entity needs to be added to the memory representation; deaccentuation can signal that the linguistic expression should be mapped onto an existing discourse entity. For example, not accenting a new entity erroneously leads a listener to search for a referent in the discourse representation (Bock & Mazzella, 1983; Terken & Nooteboom, 1987). The error occurring on the accent marking on corresponding information structure may interfere with smooth processing on the ongoing speech.

There are some common patterns of the relation between the information structure and accentuation that characterize English and Mandarin. Early study of Pierrehumbert and Hirschberg (1990) and later study of Llanos et al. (2021) have already concluded that in the context of new information focus, stressed syllables of words are marked with an H* pitch accent, while in cases of contrastive focus, stressed syllables of words receive an L + H* pitch accent. To be specific, in English, the L + H* pitch accent is employed to indicate contrastive focus in relation to different segments of a discourse, while the H* pitch accent is utilized to broadly denote various types of novel information (Morett et al., 2021; Tang et al., 2023; Wells, 2006). H* pitch accent indicating new information, typically realized with high pitch within the speaker's pitch range, denoting information focus. The contrastive pitch accent L+H* indicating the contrastive focus in sentence in relation to different segments of a discourse. It is characterized by an initial low pitch followed by a sharp rise to a high pitch on the accented syllable, signaling the contrastive focus (Llanos et al., 2021; Pierrehumbert & Hirschberg, 1990).

But it is important to notice that the relation between the information structure and accentuation is not always "one-to-one." In general, two aspects emerge. First, the relation between types of information structure and pitch accents is not a simple one-to-one correspondence. The function of the contrastive L + H* pitch accent can sometimes overlap with that of the new information H* pitch accent, which is supported by Watson et al.'s

eye-tracking study (2008). Specifically, Watson et al. (2008) used a set of instruction consist of three sentences: *Click on the camel and the dog*; *Move the dog to the right of the square*; and *Now, move the camel (H*)/candle (L+H*) below the triangle*. The first sentence creates a potential contrast set between the *camel* and *dog*. The second sentence makes one of the entity become highly accessible in the discourse (e.g. *dog*). The third sentence further creates a contrast with the potential object (e.g. *camel/candle*). In the third sentence, the given information *camel* as a contrast referent in the discourse is produced with H* pitch accent associated with new information structure, while the new information *candle* as a new referent was produced with L+H* pitch accent associated with contrast information. The results showed that an L + H* pitch accent directed listeners' look towards contrastive referents (e.g. *camel*), while an H* pitch accent prompts listeners' attention to both new and contrastive referents (e.g. *camel* and *candle*). The H* pitch accent prompts listeners' looks towards the *camel* and *candle*, which rise at roughly the same rate as the word unfolds. Therefore, based on these previous studies, pitch accent patterns that H* and L + H* are critical prosodic cues to distinguish new information focus and contrastive focus in English. However, new information focus and contrastive focus do not always neatly align with specific pitch accents; instead, the pitch accent patterns could be the same, either H* or L + H*, to indicate these two focus types.

Second, the phenomenon of *focus projection* (Baumann & Schumacher, 2012; Selkirk, 1995) in language processing, challenges the "one-to-one" relation between the accentuation according to information structure. This *focus projection* could also challenges listeners' rapid comprehension on the relation between the contrastive pitch accent and information structure. According to Selkirk's *focus projection* theory, any accented word in a sentence is "F-marked," following two rules. Firstly, F-marking of the head of a phrase, such as the noun in a noun phrase or verb in a verb phrase, allows F-marking of the entire phrase. Secondly, F-marking of an internal argument of a head permits F-marking of the head itself, for example, for a verb phrase containing a verb and an direct object, if the direct object is accented (F-marked), then the verb as the head of the verb phrase can also be F-marked, which can receive accentuation or prosodic prominence (Birch & Clifton, 1995). Therefore, affected by the focus projection phenomenon, the constituents that are

F-marked (excluding the focused word of the sentence) are interpreted as new information. In the context of focus projection, if a contrastive pitch accent is placed on a word or new information that is not the primary focus of the sentence but rather a sub-constituent within a broader focused context, it may not effectively capture listeners' attention or facilitate their processing and comprehension. In such cases, listeners may be more focused on the broader range of the sentence rather than the accented new information or word.

Therefore, the challenges to the "one-to-one" relationship between accentuation and information structure potentially undermine the previous assumption that listeners consistently expect accented new information and deaccented repeated information. Accentuation is not always indicative of introducing a new entity in the discourse representation, and the absence of accent does not always suggest seeking an existing entity in the representation to match the linguistic expression.

This also suggests that listeners may not always exhibit processing difficulties when encountering errors in accentuation marking. Birch and Clifton's study (1995) revealed that if a suitable accent occurred elsewhere within the phrase, the absence of accents on new information and the presence of accented old information were acceptable and did not result in processing difficulty. Further research in this area could provide valuable insights.

5.4.1.2 Relation between accentuation and information structure in English

English, as a non-tonal language, employs both phonological and phonetic means to achieve contrastive focus. Phonological means involve changes at higher and more abstract levels of language, such as manipulating stress patterns and deciding whether to accent words. Phonetic means involve changes at the physical level of language, encompassing adjustments in pitch, loudness, duration, articulation, and other acoustic speech sound qualities (Yang & Chen, 2018). Contrastive focus in English is encoded by a combination of pitch (fundamental frequency, F0), duration, intensity cues and vowel quality (Chrabaszcz et al., 2014; Saha & Mandal, 2018), normally applied primarily to the lexically stressed syllable of a focused word. In many varieties of English, the perceptually highest pitch point typically coincides with the onset or the start of the stressed syllable in

the intonation phrase's most emphasized word (Ladd et al., 2000; Trofimovich & Baker, 2007).

These two types of focuses in English, L+H* and H*, are showed in Figure 5.1. The Figure 5.1a illustrated the discourses (Llanos et al., 2021) like in (5) with the name *"Marianna"* as the new information produced in a H* pitch accent. The Figure 5.1b illustrated the discourse (Llanos et al., 2021) like in (6) with the name *"Marianna"* as the information contrastive to other segments in the discourse, which is accented with a L + H* pitch accent.

(5) A: Who made the marmalade?

B: **MARIANNA** made the marmalade.

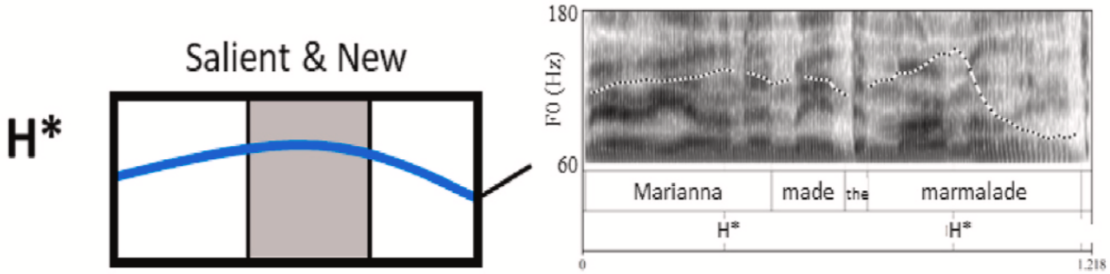
(6) A: Who made the marmalade? Marianna or you?

B: **MARIANNA** made the marmalade.

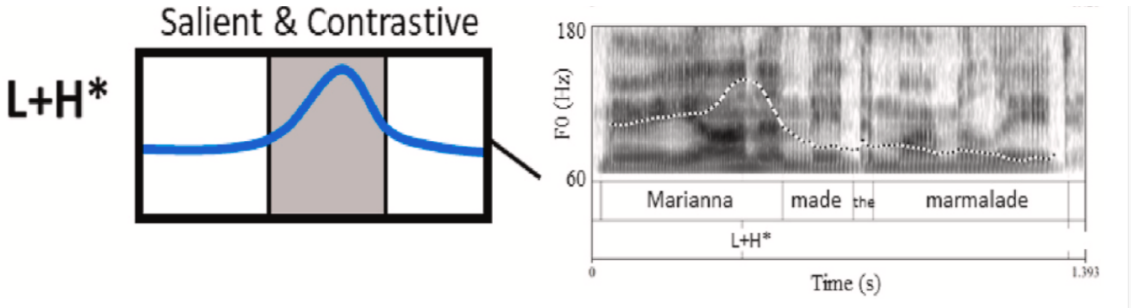
5.4.1.3 Relation between narrow focus and information structure in Mandarin

The realization of focus can vary across languages. Mandarin differs from English in how it conveys contrastive focus. Pitch in Mandarin is important to semantics recognition (Singh & Chee, 2016). As a tonal language, Mandarin utilizes a system of four lexical tones, leading to a situation where the pitch of a syllable can result in changes of its intended semantic meanings. Therefore, the use of F0 in focus is different from English. Instead of using accenting, Mandarin employs an extended pitch range and prolonged duration to indicate contrastive focus (Y.-C. Lee et al., 2016; Xu, 1999; Yang & Chen, 2018).

The experimental work in Mandarin of Chen and Braun (2006) showed that new information, has longer duration and larger F0 ranges than given information. The tone and intonation coexist in Mandarin, although they interact, the acoustic characteristics of focus in Mandarin are consistent regardless of the type of tone, where there is an elevation



(a) Pitch accent H* on *Marianna* as new information.



(b) Pitch accent L+H* on *Marianna* as contrastive information.

Figure 5.1: Reprinted from "The neural processing of pitch accents in continuous speech," by Llanos et al., 2021, *Neuropsychologia*, 158, p.107883

in the pitch range on the focused word and a subsequent reduction in the pitch range in the part of speech that follows the focus (Chen & Braun, 2006; F. Liu & Xu, 2005; Y. Liu & Ning, 2018).

The four Mandarin tones and their modulation under narrow focus are illustrated in Figure 5.2 (Y.-C. Lee et al., 2016). Applying data from English, where L + H and H* correspond to different information structural categories, directly to Mandarin is not feasible. Unlike English, where information structure is indicated by altering the *shapes* of F0 contours, Mandarin does not rely on F0 contour shapes for conveying information structure (T. Wang et al., 2020). In Mandarin, the shape of the F0 pitch contour, a major cue for lexical tone, remains unchanged when conveying information structure. There are alterations in the F0 contour shift, and increase in intensity, and pitch ranges to signify contrastive or new information (Y.-C. Lee et al., 2016; Ouyang & Kaiser, 2015).

The sentences and spectrograms provided by (Wu & Ortega-Llebaria, 2017) serve as examples illustrating the manifestation of narrow focus within a sentence. In Figure 5.3,

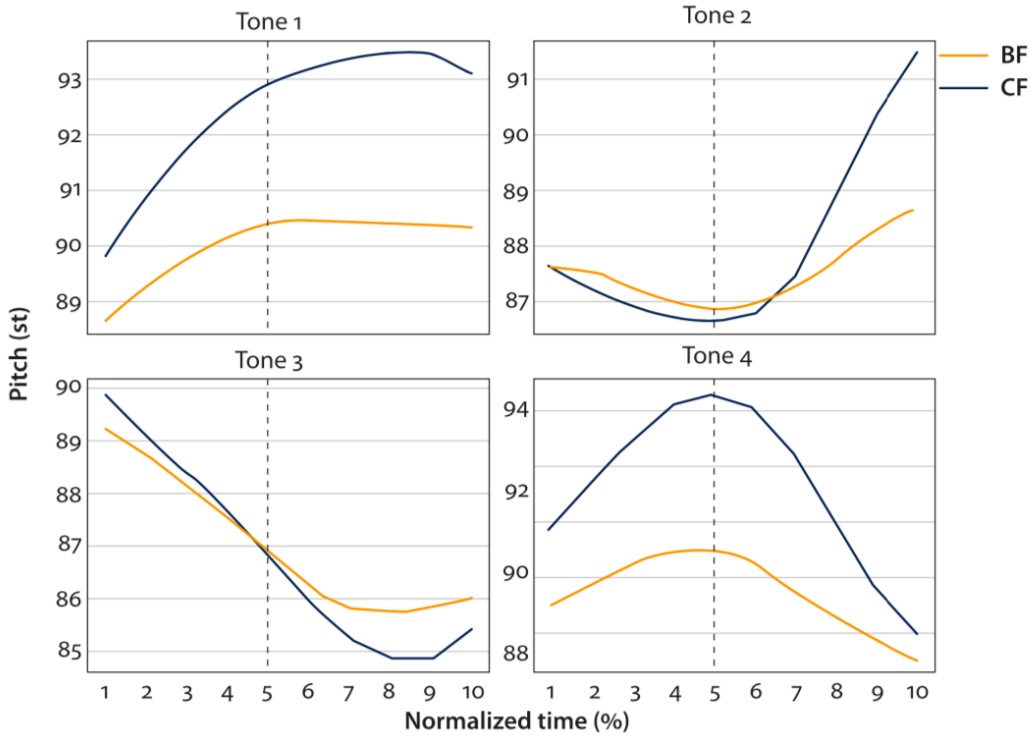


Figure 5.2: Each four tones and tones with narrow focus. BF indicates broad focus, CF indicates the corrective focus (contrastive focus). Reprinted from "Production and Perception of Tone 3 Focus in Mandarin Chinese," by Lee, Wang and Liberman, 2016, *Frontiers in Psychology*, 7:1058, p5

spectrogram (a) displays the single syllable *hua4* within a sentence with a broad focus, as extracted from the carrier sentence (8). Spectrogram (b) corresponds to the character *hua4* under narrow focus extracted from sentence (9). In spectrogram (b), there is an increase in pitch range, while maintaining unchanged F0 contour shapes.

(7) Question:

ta shuo shenme

he say what

"What did he say?"

(8) Broad focus:

hua4 hen piaoliang

Picture very beautiful

"Pictures are very beautiful."

(9) Narrow focus:

hua4 hen piaoliang

Picture very beautiful

"**Pictures** are very beautiful."

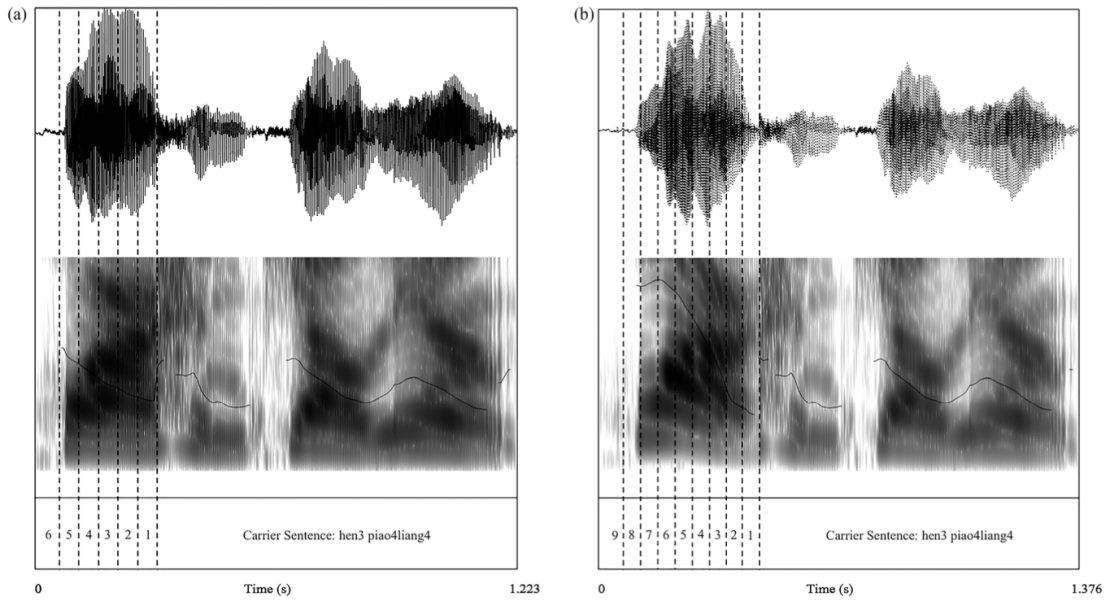


Figure 5.3: Spectrograms for single characters hua4(a) with broad focus and hua4(b) with narrow focus extracted from the carrier sentence. Reprinted from "Pitch shape modulates the time course of tone vs pitch-accent identification in Mandarin Chinese," by Wu and Ortega-Llebaria, 2017, *The Journal of the Acoustical Society of America*, 141(3), p.2267

Therefore, in contrast to English, the pitch information like F0 in Mandarin expresses both lexical meaning due to Mandarin's tonal nature and also intonation meaning used for contrastive focus encoding information structure. The prosodic cues in Mandarin can serve both lexical and discourse-level functions (Ouyang & Kaiser, 2015). The coexis-

tence of both the lexical (tone) and post-lexical function (intonation) of f0 in Mandarin has been denoted as the "multiplexing of the f0 channel" (Zhang et al., 2020). This dual role of pitch makes the mapping between the prosodic features used for contrastive focus and their phonological function less clear and direct in Mandarin than English. Especially for Mandarin L1 children before the age of 10, who found it challenging to establish an adult-like form-function mapping of contrastive focus in Mandarin (Tang et al., 2023). However, English L1 children are able to begin establishing this mapping at the age of 6 in English (K. Ito et al., 2014). Consequently, the language-specific phonological structure used to express contrastive focus in the native languages of listeners, L1 English and Mandarin, implies that Mandarin and English-Mandarin L2 listeners may employ distinct prosodic perceptual strategies when processing Mandarin (Tang et al., 2023).

For example, Liu and Ning (2018) provided evidence to support the idea of distinct prosodic perpetual strategies across Mandarin native listeners and English native listeners learning Mandarin as L2. In their study, listeners listened to a series of Mandarin sentences. Each sentence contained a disyllabic combination intended to be recognized as the focus, marked by emphasis through acoustic cues such as pitch (enlargement of pitch range), duration (longer syllable length), and intensity variations (greater intensity). Listeners were tasked with identifying the focus in these Mandarin sentences based on the emphasis marked by the acoustic cues. This task assessed their sensitivity to pitch, duration, and intensity cues in recognizing focus. The results showed that English speakers, when compared to native Mandarin speakers, showed a different pattern in identifying focus. While Mandarin speakers' perception of focus was significantly influenced by pitch variation, which is a primary cue in Mandarin's tonal system, English speakers' performance indicated that they were more influenced by duration and intensity cues. This difference suggests a reliance on prosodic features more prominently used in their native language (English) for signaling stress or emphasis, rather than the pitch variations more characteristic of tonal languages like Mandarin (Y. Liu & Ning, 2018). Therefore, this suggested the impact of native language phonological and prosodic structures on the perception of focus in L2. Mandarin speakers naturally prioritize pitch due to their language's tonal nature, while English speakers focus on duration and intensity, reflecting

the prosodic characteristics of their native language. The prosodic perceptual strategies adopted by listeners is under the influence of these inherent perceptual biases.

5.4.2 Native listener's ability in using pitch accent as prosodic cue

For the current project, the role of pitch accent as a prosodic cue is investigated to understand both native and non-native listeners' perceptual adaptability to various cues. A key area of focus is how listeners navigate the alignment or disparity between the prosodic system and the information structure. This central focus gives rise to two critical aspects regarding non-native speech processing. Firstly, listeners' ability to concentrate on the target noun marked by the pitch accent, a phenomenon known as the "search for focus" effect (Akker & Cutler, 2003) (reviewed in *Section 5.1*), which highlights a specific listener behavior where the pitch accent on a target noun draws concentrated attention. This "search for focus" contributes to the swift identification and comprehension of semantic structure within spoken discourse. Secondly, listeners' ability in utilizing pitch accent to establish and anticipate the information structure of the forthcoming noun. The anticipatory use of pitch accent enables listeners to make predictions about whether an upcoming noun will introduce new information or refer back to something already mentioned within the discourse.

Concerning the first aspect, native listeners demonstrate the "search for focus" effect (Akker & Cutler, 2003). As previously illustrated, native listeners exhibit high sensitivity to the pitch accent on the target word and can process sentence accent patterns, along with the discourse implications they convey, with efficiency and rapidity: they detected phoneme targets in sentences more rapidly when the target-bearing words are in an accented or focused position (Akker & Cutler, 2003; Ip & Cutler, 2020). Therefore, it is confirmed that accented or focused targets facilitate the native listeners' detection of the target. It is plausible that the contrastive pitch accent or emphasis placed on the target word renders it more salient and perceptibly louder, thereby facilitating its detection by listeners.

Similarly, previous studies have also revealed a phenomenon known as "depth of process-

ing" (Brunellière et al., 2019; Li & Lu, 2011; Li & Ren, 2012; L. Wang et al., 2011), demonstrating that native listeners can effectively integrate pitch accent information and semantic information at the target noun. For example, Wang et al.'s ERP study (2011) investigates the impact of information structure, specifically focus and pitch accent, on the depth of semantic processing, as reflected in the N400 effect (a significant N400 as an indicator of semantic processing). They employed native Dutch listeners as participants. Listeners listened to sets of "wh-question" and answer sentences in four conditions based on three manipulated factors: context (whether the critical word is in focus or non-focus based on the question), semantic congruency (whether the answer is semantically congruent or incongruent with the question), and prosodic accentuation (whether the critical word is accented or unaccented): focus and accented (*"What kind of vegetable did mum buy for dinner today? Today mum bought EGGPLANT/BEEF for dinner"*); focus and unaccented (*"What kind of vegetable did mum buy for dinner today? Today MUM bought eggplant/beef for dinner"*); non-focus and accented (*"Who bought the vegetable for dinner today? Today mum bought EGGPLANT/BEEF for dinner"*); non-focus and unaccented (*"Who bought the vegetable for dinner today? Today MUM bought eggplant/beef for dinner"*). The study found that the semantic processing of nouns was enhanced when these target nouns were emphasized by a prosodic pitch accent. When the focal target noun is accented, and this noun is inappropriate to be provided as the answer in a question-answer pair, such as in the question-answer sentence *"What kind of vegetable did mum buy for dinner today? Today mum bought BEEF for dinner"*, the noun *"beef"* is the focus and inappropriate in the question-answer pair, also with pitch accent, this condition is associated with the most pronounced N400 response compared to other conditions. This study demonstrated that the interaction between accentuation and context critically influences the depth of semantic processing, particularly highlighting that deeper processing is elicited when both accentuation and context converge on the same, semantically incongruent information. Conversely, when there is a mismatch between accentuation and context, or when the focused information is semantically congruent, the depth of semantic processing is diminished, as evidenced by smaller N400 effects.

Similarly, in Brunellière et al.'s study (2019), listeners heard a set of French sentences

where the sentence-final target word was either accented or neutral, and semantically congruent or incongruent with the preceding context, such as *Créer des bonbons : ils ont un beau métier, les confiseurs/les pisteurs* ("Creating sweets, they have a great job, confectioners/the members of sky patrol"). In this example, *les confiseurs* ("confectioners"; semantic congruent) and *les pisteurs* ("the member of sky patrol"; semantic incongruent) were either deaccented or accented. Listeners were then asked to perform a lexical recognition task, where they had to quickly and accurately indicate whether a set of presented words had been used as the final words in the sentences they heard. The results showed that an incongruous final word with emphasis (e.g. *les pisteurs*, "ski patrol members") caused a semantic anomaly, eliciting a larger N400, thus illustrating how prosody deepens semantic processing. Similarly, previous studies have shown that semantically incongruent words elicit a significant N400 effect in both accented and greatly accented conditions, but not in the de-accented condition (Li & Ren, 2012). Additionally, accentuation facilitated word memory—words receiving pitch accent were better retained in short-term memory, and accentuation enhanced the semantic relatedness between the target word and the preceding word (Li & Lu, 2011).

In addition to the contrastive pitch accent on the target noun itself to enhance the search for focus and deepen the processing, the presence of a contrastive pitch accent on the adjective preceding the target noun can also facilitate the detection of the target noun. Weber et al.'s eye-tracking study (2006) demonstrated that native German listeners could effectively use prosodic cues, specifically contrastive pitch accents, to anticipate and identify referents in spoken language before the referent noun was pronounced. In the study, when the adjective *"red"* in the phrase *"RED scissors"* received a contrastive pitch accent, participants were directed to select the image of red scissors from a display that also featured purple scissors. Listeners looked at the intended picture of *"RED scissors"*, which had a contrastive pitch accent on the adjective, more rapidly than when the adjective was deaccented. This demonstrated a heightened attentional focus and quicker referent determination in contrast to conditions where the adjective was not emphasized with a pitch accent. The effect underscores the significant role of contrastive pitch accents in enhancing the salience of specific attributes (in this case, color) during auditory process-

ing, thereby facilitating the rapid localization of the target object among a set of potential referents, even before the noun was fully articulated. Therefore, it was showed that the listeners were able to utilize the prosodic information to facilitate rapid and anticipatory reference resolution in spoken discourse comprehension.

Pitch accent serves as a significant marker of information structure status. In the context of native listeners, early research on prosodic cues highlighted the role of contrastive pitch accents in facilitating recognition. The accented words are processed more rapidly compared to unaccented ones (Cutler, 1976; Cutler & Foss, 1977). For example, an early study of Cutler (1976) has demonstrated that accented words were processed more quickly than unaccented ones, as evidenced by shorter response times in detecting the initial phoneme of accented words. This observation suggests that accentuation accelerates the recognition of new information conveyed by accented words. Subsequent research (Most & Saltz, 1979; Nooteboom & Kruyt, 1984, 1987; Terken & Nooteboom, 1987) further confirmed and contributed to the widespread recognition of the fact that when the word conveying given information lacked accent, its identification and comprehension occurred more swiftly, while new information was identified and comprehended more rapidly when the word expressing it was accented.

Research conducted across different languages has consistently demonstrated that prosodic cues play a crucial role in conveying information structure for listeners and native listeners have showed high ability in utilizing this prosodic cue to determine the information structure of the upcoming noun, a phenomenon widely acknowledged in various linguistic contexts. Previous neurolinguistic studies examining the impact of prosodic cues on utterance information structure have revealed that native listeners possess a sensitivity to these cues, allowing them to predict the information structure conveyed in utterances. For instance, in a study by Baumann and Schumacher (2012), the N400 component was utilized to investigate how native German listeners utilize prosodic cues to distinguish information structure. The study manipulated accentuation patterns in the stimuli to indicate different information structures. Their findings indicated that the N400 component responds sensitively to variations in information structure and associated prosodic cues. Specifically,

inappropriate accentuation, such as the absence of accents on new items or the presence of unnecessary accents on given items, resulted in discernible N400 modulations, suggesting conflicts between the prosodic cue and information structure.

Native listeners' proficiency in employing pitch accent to ascertain the information structure of forthcoming nouns also extends to their capability to utilize pitch accent for reference resolution during spoken language comprehension. These processes both necessitate listeners to interpret and predict the discourse structure through the utilization of pitch accent. Native listeners were showed to be able to use pitch accent to resolve references in spoken language comprehension. Reference resolution refers to the process of determining the intended referent of a pronoun or a noun phrase in discourse. In reference resolution, native listeners interpreted an accented word as a newly-introduced discourse entity while the deaccented words as an anaphoric. Therefore, prosody can provide important cues, such as pitch accents or phrasing, which helps listeners identify the referential relationships between different parts of a sentence or discourse (Dahan et al., 2002). Dahan et al.'s eye-tracking study (2002) asked listeners to move the picture to a position mentioned in instructions in the form of two sentences. The first sentence asked listeners to move an object below a shape: *"Put the candle/candy below the triangle"*. The second sentence contained an accented or deaccented definite noun phrase which referred to the same object or introduced a new entity: *"Now put the candle ABOVE THE SQUARE"* or *"Now put the CANDLE above the square"*. The results showed that fixations to the competitor (e.g. "candy") demonstrated a bias to interpret deaccented nouns as anaphoric (repeated information) and accented nouns as nonanaphoric (new information). When the underlined target word was accented, at the beginning of the accented noun in the second sentence, listeners were more likely to fixate new pictures which were not previously mentioned in the first instruction. By contrast, if the target word was deaccented, listeners were more likely to fixate on old pictures that were previously mentioned in the first sentence. Therefore, native listeners are able to utilize the prosodic cue, such as the accentuation to interpret the referent in sentences, and have a bias towards a certain interpretation according to the variations of the accentuation.

5.4.3 Non-native listener's ability in using pitch accent as prosodic cue

Non-native listeners, much like their native counterparts, demonstrate the ability to discern and use prosodic cues in L2 sentence processing and comprehension to some extent, indicating that they can exploit prosodic structure in their L2. Although the studies discussed here showed that non-native listeners exhibited a delay and difficulty in discerning and utilizing prosody compared to native listeners, they still display the capacity to use prosodic cues and integrate information across various domains to construct information structures when perceiving non-native language speech (Akker & Cutler, 2003; Perdomo & Kaan, 2021).

Fundamentally, non-native listeners have been demonstrated to utilize fundamental prosodic features such as F0 movements and vowel duration to assess, predict, and detect intonational prominence in sentences (Rosenberg et al., 2010; Wagner, 2005). Moreover, non-native listeners also demonstrate the ability to utilize prosodic cues for various linguistic purposes. For pragmatic purpose, research indicates that non-native listeners can use prosodic elements such as timing and pitch to evaluate the naturalness of a speaker's pronunciation (Tsurutani et al., 2010). For syntactic processing purpose, non-native listeners have been observed to employ prosodic cues to disambiguate syntactic ambiguity while processing spoken languages (O'Brien et al., 2014). In the domain of segmentation, non-native listeners also exhibit the capability to segment speech, regardless of their proficiency in the target L2. They use different prosodic cues, such as duration rather than F0 rise employed by native listeners, to determine word boundaries (Coughlin & Tremblay, 2012; Pintér et al., 2014; Sanders et al., 2002; Tremblay et al., 2012, 2018). However, the use of prosodic cues for segmentation may be delayed in word recognition compared to native listeners (Coughlin & Tremblay, 2012) and is influenced by the non-native listeners' knowledge of their native language (Hanulíková et al., 2011; Tremblay et al., 2018; Weber & Cutler, 2006). Beyond these functions, non-native listeners can predict upcoming nouns based on contrastive pitch accents on adjectives preceding the nouns (Foltz, 2021). This implies that non-native listeners can develop expectations about whether or not a noun will be repeated, depending on consistent prosodic patterns, such as sentences in which the adjective carries a contrastive pitch accent consistently lead to a repeated

noun (e.g. *"Click on the red duck. Click on the GREEN..."*). The information structure of a noun is based on the presence of a contrastive pitch accent on the preceding adjective. Nevertheless, the predictive processing of non-native listeners exhibits a noticeable delay when compared to that of native listeners. These previous studies highlight that, despite a reduced ability and influence from native languages, non-native listeners still have the ability, to some extent, to utilize prosodic cues as suprasegmental indicators for various speech processing purposes.

Similar to the discussion in the preceding section (*Section 5.4.4*) for listeners in the native speech processing, two aspects remain important for non-native listeners regarding the utilization of pitch accent as a prosodic cue: firstly, their ability to identify the target word based on the pitch accent, and secondly, their proficiency in using pitch accent to predict the information structure of the upcoming noun.

In relation to the first aspect, emphasizing a word leads to both prolonged word duration and enhanced spectral clarity. Consequently, accented syllables are often acoustically more salient, making them easier to be detected and processed at early stages. This prompts listeners to actively focus on segments of speech where accent is anticipated. However, the previous study (Perdomo & Kaan, 2021) has argued that the performance of non-native listeners in this proficiency of utilizing prosodic cues for sentence accent detection is inconsistent, primarily depending on characteristics of their native language's prosodic cues, proficiency in the target non-native languages, and the listening environment. As an illustration, in a phoneme monitoring study by Scharenborg et al. (2020), stimuli were played in a noisy background to investigate whether pitch accent on the target word aids speech perception in the presence of background noise. English non-native listeners (Dutch, Finnish, and French non-native listeners of English) were asked to listen for specific initial target phonemes (e.g. /k/) in the target word within sentences. The example sentence with accented target was like *"The remains of the CAMP were found by the tiger hunter"*, while the example sentence with deaccented target word was like *"The remains of the camp were found by the TIGER hunter"*. The findings indicated that both native and non-native listeners were more accurate in detecting a target phoneme in the accented

target word compared to the deaccented word. Therefore, non-native English listeners detected demonstrated a proficiency in utilizing prosodic cues to signal sentence accent for target sound detection that was comparable to that of native English speakers. This suggests a notable capability among non-native listeners to effectively leverage prosodic information in speech perception to "search for focus." Moreover, the pitch accent on the target facilitates listeners' ability in searching for focus. This finding underscores the role of pitch accent as a useful cue for listeners, both native and non-native, in processing and understanding speech, particularly in contexts where discerning the focus of a sentence is critical for comprehension.

However, as is common in many studies on non-native speech processing, the results regarding listeners' performance exhibited inconsistency. In some studies, non-native listeners were unable to achieve the same level of proficiency as native listeners in real-time speech perception, processing and comprehension. Conversely, other studies found that non-native listeners performed comparably to native listeners. This suggests that although native listeners generally outperform non-native listeners in speech processing and comprehension, it is not that simple. For example, Akker and Cutler's monitoring experiment (2003, as summarized in Section 5.1.4), revealed that, unlike native listeners, non-native listeners did not exhibit the same level of proficiency in utilizing prosodic cues to anticipate the location of accent in a sentence or react significantly faster to accented targets than to deaccented targets (Akker & Cutler, 2003). However, the delay in processing prosody has not been consistent across previous studies. While some studies have revealed a delayed processing of prosodic information for non-native listeners, others have not. For instance, Scharenborg et al. (2020) did not find evidence of non-native listeners processing prosodic cues with a delay, and the groups of non-native listeners demonstrated comparable performance to native listeners in detecting focus by utilizing pitch accent on the target.

Scharenborg et al. (2020) provided insights into the factors influencing the performance of non-native listeners. They concluded that, similar to native listeners, non-native listeners can effectively utilize accentuation on the target word to aid in word detection, but

non-native listeners' performance in L2 real-time sentence processing could depends on their L1. Dutch and English are regarded as two languages that exhibit minimal mismatches at the phonological level, with only a relatively small discrepancy at the level of individual sounds. This is because the prosodic cues and structures for sentence accent and prosodic processing in Dutch and English are highly comparable (Akker & Cutler, 2003; Scharenborg et al., 2020). However, in their experiment, French native listeners learning English as their L2 performed comparably to English native listeners, despite differences in prosodic patterns between French and English. Conversely, Dutch native listeners learning English as their L2 exhibited a different pattern. Therefore, Scharenborg et al. (2020) concluded that the ability to utilize sentence accent for speech perception is not solely determined by the relative similarity between prosodic cues in native and non-native languages. When non-native listeners encounter challenges in processing the prosody of non-native English, such as rhythm, intonation, and stress patterns, they tend to rely on familiar prosodic patterns from their native language. Moreover, as summarized in Section 5.1.3 *Cue weighting theory*, Scharenborg et al. (2020) also pointed out that proficiency does not straightforwardly predict sentence accent perception performance. Recall that listeners who prioritize distributed prosodic cues tend to perform better in non-native prosodic perception, while listeners who depend more on local cues than on distributed cues, are more susceptible to difficulties in perceiving prosodic cues.

Regarding the second aspect of the central consideration in utilizing pitch accent as a prosodic cue in relation with information structure for non-native listeners, it is observed that non-native listeners can employ it to construct information structure and contrast sets. However, their ability to predict upcoming referents in spoken language is diminished compared to that of native listeners (Foltz, 2021; Perdomo & Kaan, 2021). In Perdomo and Kaan's (2021) eye-tracking experiment involving native Mandarin listeners learning English as their L2, the focus was on how contrastive pitch accent aids in building information structure and to what extent this prosodic cue can be used predictively. Participants listened to English sentences like *"We ate Angela's ice cream/cake but saved Benjamin's/BENjamin's cake in the fridge"*. The researchers employed contrastive pitch accent on the second proper noun (e.g. *"BENjamin's"*) to signal similarity to the ear-

lier mentioned referent, expecting that when the accent is on Benjamin's name, listeners would direct their gaze more towards the previously mentioned object. The results revealed that native Mandarin listeners successfully utilized prosodic cues, like contrastive pitch accent, and integrated prosodic and lexico-semantic information to build information structure, mirroring the performance of native English listeners. However, native Mandarin listeners exhibited a delayed effect in processing contrastive pitch accent, indicating a lack of strong evidence for predictive use of prosody compared to native English listeners. Perdomo and Kaan (2021) suggested that the reduced ability of non-native listeners could be attributed to factors such as the complexity of processing for non-native listeners and differences in marking contrastive focus in Mandarin. Notably, they found that working memory had no significant effect on non-native listeners, and the relationship between proficiency and the use of prosody was challenging to interpret, despite Mandarin native listeners with higher proficiency in English showing more looks to the target in a specific time bin, suggesting faster recognition of the target word.

The study by Takahashi et al. (2018) suggested that non-native listeners exhibit sensitivity to modulations of L2 prosody but may encounter challenges in integrating this information with other sources. In their study involving Mandarin native listeners learning English as a L2, participants were presented with a behavioral task where they had to select the item on the screen corresponding to the English sentences they heard. For instance, sentences such as *"Click on the purple mittens. Now click on the SCARLET mittens"*, where the prosodic focus (L+H*) on the adjective was felicitous due to the color contrast between purple and scarlet, or *"Click on the scarlet necklace. Now click on the SCARLET mittens"*, where the prosodic focus on the adjective was infelicitous as the contrast was between the necklace and mittens. Takahashi et al. hypothesized that Mandarin native listeners would demonstrate a similar pattern to native speakers, where felicitous contrastive prosody (L+H*) on the adjective would lead to faster responses. The findings indicated that Mandarin native listeners generally responded more quickly in the felicitous condition compared to the infelicitous one, resembling English speakers, although the effect was smaller in the Mandarin speaker group. This suggests that Mandarin native listeners, while exhibiting a reduced ability, were still capable of utilizing prosodic cues to establish

information structure, similar to English native listeners. The contrastive pitch accent on the adjective indicated an old information mentioned earlier.

The ability of listeners to use prosodic cues, including pitch accent, vowel lengthening and pitch movement, for understanding speech varies with their linguistic background. Native listeners have showed proficiency in using these cues for segmenting continuous speech and identifying referential targets within discourse, with their performance influenced by the specific characteristics of their native language (Tyler & Cutler, 2009). Tyler and Cutler (2009) investigated how native listeners of English, French, and Dutch utilized vowel lengthening and pitch movement cues for continuous-speech segmentation. Their findings revealed a dual pattern: the ability of utilizing prosodic cues for speech segmentation was universal among native listeners, however, each language group utilized these cues in a different way, revealing language-specific patterns. These three groups of native listeners all benefited from the vowel lengthening cue for word segmentation when it was placed at the right-edge position of words, in which the final vowel sound of a word is extended and lengthened, indicating a boundary or the end of a word. As for the pitch movement cue to indicate prominence in the sequences, English native listeners employed it at the onset positions, while French listeners utilized it at the terminal position, and Dutch listeners made use of it at both the onset and terminal positions. This indicates that English listeners rely on pitch prominence at the beginning of sequences to identify the beginning and initial syllable of content words within an utterance. For French listeners, they rely on pitch prominence cues that are typically associated with the right-edge of sentences to mark the ending. In Dutch, these cues can be found at both the beginning and the end of sentences or phrases, suggesting a more flexible use of pitch prominence cues by Dutch native listeners to identify both sentence beginnings and endings.

For non-native listeners, the strategies employed in utilizing prosodic cues are often influenced by their native language, leading to variations in their ability to leverage these cues effectively. This diversity in prosodic cue utilization among listeners from different linguistic backgrounds can result in distinct performances in tasks involving prosody (Scharenborg et al., 2020).

Although non-native listeners were capable of using prosodic cues to some extent, previous research has indicated that non-native listeners experienced a significant difficulty in utilizing these prosodic cues compared to native listeners (Foltz, 2021; Perdomo & Kaan, 2021; Vanlancker–Sidtis, 2003). For example, as prosody is an important cue to provide semantic interpretations of sentences, in Vanlancker-Sidtis's study (2003), sentences such as *She had him eating out of her head* was produced by two speakers with a literal and idiomatic meaning intended for listeners to interpret. The results showed that non-native listeners were not as good as native listeners at utilizing prosodic information to judge if a sentence is idiomatic or literal. Therefore, non-native listeners may not adequately recognize the prosodic patterns that signal idiomatic usage and they might default to a more literal interpretation of sentences. Pennington and Ellis (2000) provided listeners with a list of sentences that included prosodic cues, such as *Is he/He driving the bus?* (with or without special emphasis on *HE*), and *He's a good boy, isn't he?* (with falling or rising intonation). Listeners were instructed to listen carefully and, after a short break, were given a new list of sentences to determine whether or not each sentence matched one from the previous task. The results showed that non-native listeners had poor recognition and memory for prosodically signaled information compared to native listeners. In the second experiment of Pennington and Ellis' study, listeners were deliberately directed to pay attention to intonation and prosodic differences. While there was some improvement in recognizing marked prosodic focus, non-native listeners overall still demonstrated limited ability to use prosodic cues effectively.

Previous studies also showed that non-native listeners prefer prosody less than other speech cues (e.g. contextual cue, lexical-semantics) (Clahsen & Felser, 2006b; Juffs & Harrington, 1995; Vanlancker–Sidtis, 2003; Ying, 1996) and were less able to integrate the prosodic information with semantic information (Akker & Cutler, 2003; Ganga et al., 2024). For example, in Ying's study (1996), L2 English learners read a sentence with neither additional prosodic nor contextual cues, such as *The girl saw the man with a special pair of glasses*, and they showed a preference to attach the prepositional phrase *with a special pair of glasses* to the verb phrase *saw*. Then L2 English learners listened to a sentence carrying an additional prosodic cue, such as *The man talked [prosodic break] to the*

girl with a sense of humor, in which a short pause was inserted between the verb *talked* and the preposition *to*, with an unbroken intonation contour for the rest of the sentence, indicating that the noun phrase *the girl* and the prepositional phrase *with a sense of humor* is a single unit. L2 English learners were also provided with a sentence with contextual cues, such as *There were two girls. One of them had a sense of humor, and the other did not. The man talked to the girl with a sense of humor*, in which the contextual information biased the attachment of the prepositional phrase *with a sense of humor* to the noun phrase *the girl*. The results showed that the prosodic information assisted little with non-native listeners. L2 English learners heard the sentence with a prosodic cue, but there was not much difference compared to when they processed the sentence without prosodic break information. However, they showed a significant preference to interpret the prepositional phrase and noun phrase as a single unit when provided with contextual information. This suggested that contextual information may have greater impact on non-native listeners than prosodic information in sentence processing.

As discussed previously, Akker and Cutler's study (2003) showed that non-native listeners were able to search for and identify the focus in the sentence based on prosodic information. However, Akker and Cutler also provided the contextual cue in the contextual questions such as *Which bones were found by the archaeologist?* and *Which archaeologist found the bones?*, and found that non-native listeners relied more heavily on the contextual cue than the pitch accent used in the target sentence to expect and determine the focus of the upcoming target sentence. Akker and Cutler (2003) commented that although non-native listeners can exploit prosodic cue to some extent in their L2, but they are more likely to semantic-driven, in other words, highly context-dependent, rather than focusing on prosodic features during sentence processing.

The study by Ganga et al. (2024) showed that non-native listeners had difficulties integrating prosodic information (pitch accent) with semantic information when processing English sentences with *only*. Ganga et al. (2024) provided object-focus sentences such as *John is only throwing the BALL, not throwing the pen* and verb-focus sentences such as *John is only THROWING the ball, not kicking the ball*. Listeners had to process these sen-

tences and their prosodic cues to determine and understand where the focus was placed. Dutch speakers learning English as a second language exhibited less pronounced and delayed ERP effects compared to native listeners for processing prosodic cues in both object-focus and verb-focus sentences, suggesting that they did not process the prosodic emphasis on the noun or verb as effectively as native listeners. This also indicates that non-native listeners had difficulty integrating the prosodic information (pitch accent) with the semantic content of the sentence.

Therefore, previous studies have showed that non-native listeners can use and integrate prosodic information to some extent, similar to native listeners. However, they also experience certain deficits in utilizing and integrating prosodic information during sentence processing. As demonstrated in previous research, non-native listeners not only exhibit delays and insensitivity in processing and memorizing prosodic information but also rely more heavily on contextual and lexical semantic information than on prosodic information for interpreting sentences, processing sentence focus and anticipating upcoming targets, compared to native listeners.

5.5 Summary

As outlined in the preceding sections, both native and non-native listeners use prosodic and semantic information, but not to the same extent. Compared to native listeners, non-native listeners typically experience delayed responses and detection of targets with semantic or prosodic information in sentence processing (Coughlin & Tremblay, 2012; Dijkgraaf et al., 2017; Foltz, 2021; A. Ito et al., 2018; Perdomo & Kaan, 2021). Moreover, non-native listeners have also been showed to be more driven by lexical-semantic cues in sentence processing (Clahsen & Felser, 2006b; Juffs & Harrington, 1995). Under certain circumstances, non-native listeners are even inclined to abandon the use of prosodic cues. For example, studies by Akker and Cutler (2003) and Ip and Cutler (2020) suggest that non-native listeners' inability to utilize and integrate prosodic information can be attributed to the additional functional load it creates for them (Ip & Cutler, 2020; Reed & Michaud, 2015). Therefore, the previous studies showed that contextual and lexical-semantic cues are more robust cues than prosodic and syntactic cues for non-native listen-

ers (Clahsen & Felser, 2006b; Juffs & Harrington, 1995; Vanlancker–Sidtis, 2003; Ying, 1996). However, previous studies seem to lack consistency in their findings regarding non-native listeners' ability to utilize multiple cues and prioritize one over others. Additionally, details about the circumstances under which listeners' ability to use and integrate these cues varies are still unclear.

Therefore, to investigate the distinctions between native and non-native listeners in their attention weight to multiple cues, as manifested by their ability to utilize and integrate cues, the current eye-tracking experiments applied multiple cues in its stimuli. It includes pitch accent as a prosodic cue, and verb semantics and information structure as word- and sentence-level information. This part of the study poses two general questions when processing with multiple cues: (1) whether non-native listeners are capable of utilizing various cues, and (2) whether non-native listeners are less proficient at effectively integrating semantic and prosodic information interactively compared to native listeners.

Chapter 6

Eye-tracking experiment method: native and non-native listener's attention weight on cues

6.1 Experiment overview

This experiment investigates whether a consistent attentional preference exists among native and non-native listeners on certain cues during real-time speech perception, and whether this attentional preference affects listeners' ability to flexibly integrate those cues during speech perception. Therefore, the study not only examines the types of cues that capture listeners' attention but also the ability to integrate multiple cues observed within groups of native and non-native listeners.

A visual-world eye-tracking experiment was designed and conducted. In this experiment, four visual-world pictures were displayed on the screen, and participants were instructed to click on the picture corresponding to the sentence they were listening to. The auditory stimuli were presented with manipulations involving three types of language cues: pitch accent, information structure, and verb semantics.

Additional to these three types of language cues as variables in the current eye-tracking experiment, the fourth variable is the native language of participant groups. Therefore, the 4 independent variables in the eye-tracking experiment are: (1) Accent (*contrastive pitch accent* versus *neutral accent*), (2)Information structure (*old information* versus *new information*), (3) Semantics (*appropriate verb semantics* versus *inappropriate verb se-*

mantics), and (4) L1 (*English* versus *Mandarin*).

Information structure is the cue at discourse level, which is linked with pitch accent. Verb semantics evaluated listeners' attention to semantic information (referred as *effect of Semantics* in the experiment). The three language cues directed listeners' attention during the real-time sentence processing. This experiment employed the combination of pitch accent and information structure to test listeners' attention to prosodic information (referred as *effect of Accent* in the experiment). In various languages, it is widely acknowledged that prosodic structure conveys information at the level of discourse (Ouyang & Kaiser, 2015; Watson, 2010): the expected pitch accent differs depending on whether a target word is old or new information.

Two eye-tracking experiments are designed in this thesis: one with English stimuli (Experiment 3) and the other with Mandarin stimuli (Experiment 4), involving different participant groups. The designs of experiments 3 and 4 are otherwise identical.

During the experiments, listeners completed a total of 80 trials with 40 critical trials and 40 foils. In each trial, listeners heard two sentences and were instructed to follow the description to look at and click on the object mentioned in the second of the two sentences presented in a visual display. As an example trial, I used sentences from the English stimuli; for further details regarding the stimuli, please refer to the *Material* section for both the English and Mandarin stimuli parts of the experiment.

The first sentence introduced an object, for example, *"Here is a cabbage"*. The second sentence then contained the description of an action involving the target object, followed by an adverbial time phrase as in *"Susan is going to shred the cabbage after school"*. On the screen, participants saw four pictures, and were asked to identify the object named in the second sentence.

As there were three independent variables, each with two conditions, this experiment employed a 2*2*2 factorial design, resulting in a total of 8 different conditions for each target. In general, the object in the first phrase could be either the same or different from

the target in the second phrase. The target in the second phrase could be either accented with contrastive pitch accent or deaccented with neutral accent. The verb in the second phrase could either have appropriate semantics or inappropriate semantics when used with the target word.

6.1.1 Research questions

Chapter 5 reviewed evidence suggesting that non-native listeners make differential use of semantic cues compared to prosodic cues in real-time speech perception. Native and non-native listeners might give different attention weight to different cues. Therefore, these two experiments are designed to answer the following two research questions:

Question 1 Do non-native listeners use more verb semantic information than prosodic pitch accent information?

Question 2: Are non-native listeners less able to combine semantic and prosodic pitch accent information interactively compared to native listeners?

6.1.2 Predictions

In the studies reviewed in Chapters 5, native listeners demonstrated their robust ability to effectively utilize multiple cues at both the suprasegmental and discourse levels, such as semantic and prosodic cues, during real-time speech processing. Moreover, native listeners also showed proficiency in integrating multiple cues, as they have knowledge of which types of cues can coexist and which types tend to conflict with each other during real-time speech processing. Therefore, in Experiment 3 and 4, I predict that non-native listeners could have reduced perceptual flexibility compared to native listeners, as non-native listeners may be less effective in utilizing and integrating various cues. For the utilization of cues, non-native listeners might be expected to consistently rely on a certain cue, such as verb semantic information. For the integration of multiple cues, non-native listeners might also be less able to flexibly integrate different cues together compared to native listeners.

If non-native listeners are less attentive to prosody than semantic information, we expect that non-native listeners will show reduced effect of pitch accent in their ability to identify a target noun, relative to native listeners. We also expect the effect of prosody in non-native listeners to be smaller than the effect of semantic information. Finally, if non-native listeners are less able to combine cues interactively, then I expect that the interaction between Semantics and Accent could be absent: the effect of Semantics will be smaller, regardless of pitch accent. By contrast, native listeners are predicted to show the interaction: semantic processing should increase when target nouns are accented, especially when that accent correctly signals new information structure in the target noun.

To specify, the predictions are as follows:

Prediction 1 A reduced effect of Accent in non-native listeners to identify a target noun is expected when compared to that of native listeners. Within the group non-native listeners, the effect of Accent is also expected to be smaller than the effect of Semantics.

Prediction 2 No interaction between prosodic accent and verb semantics should be observed for non-native listeners, whereas the interaction is expected to be observed in native listeners. Native listeners are expected to exhibit an interaction of multiple cues, specifically an increased focus on semantic processing when target nouns are accented compared to when they are unaccented, especially in sentences containing new information.

6.2 Experiment methods

6.2.1 Participants

6.2.1.1 Experiment 3

Participants in Experiment 3 were the in-lab participants from Experiment 1. These participants voluntarily chose to continue and complete the eye-tracking experiment in the lab. To avoid fatigue, participants were given a 15-minute break before the trials began. This time included an introduction to the eye-tracking experiment, the setup of the eye-tracker, and the completion of eye-tracker calibration and validation. All participants were

recruited in the UK and completed the eye-tracking experiment in the lab setting at the University of Glasgow. All the participants were University students of the age above 18, without hearing impairment. The eye-tracking experiment for each participant lasted approximately 35 minutes, and participants were financially compensated £10 for their time. Participants included 40 English native speakers and 40 Mandarin native speakers learning English as L2. The Mandarin listeners were all Chinese students studying abroad at the University of Glasgow, used English daily with at least intermediate proficiency (IELTS score above 5.5).

6.2.1.2 Experiment 4

Experiment 3 and 4 employed different participant groups. Participants in Experiment 4 were the in-lab participants from Experiment 2. As in Experiment 3, participants were given a 15-minute break prior to the trials for introduction, setup, and calibration. All participants were recruited in the UK and completed the eye-tracking experiment in the lab at the University of Glasgow. Participants in this experiment were also university students of the age above 18, with no hearing impairments. The experiment lasted 35 minutes for each participant, and they were financially compensated £10 for their time. Participants comprised 40 native speakers of Mandarin, and 27 English native speakers learning Mandarin as L2. The English listeners had at least intermediate proficiency and had completed the intermediate Mandarin lessons at the University of Glasgow or had been exposed to a Mandarin multilingual environment for at least 2 years from an English-speaking household (family with English as a home language).

6.2.2 Materials

In both Experiment 3 and Experiment 4, each trial consists of two sentences. The first sentence serves as the introduction, while the second sentence contains a verb and a target noun followed by adverbial phrases. The inclusion of adverbial phrases and adverbial time phrases in the stimuli is aimed at avoiding having the target word at the end of sentences, preventing listeners from noticing the target due to sentence-final patterns.

All sentences were recorded by a phonetically-trained native speaker of American English

and a native speaker of Mandarin (Standard Putonghua) in a soundproof booth. In each trial, four pictures corresponding to 4 objects were displayed on the screen: a target noun, a competitor, and two distractors.

See Appendix B for a full list of English and Mandarin sentence materials showing each item in all 8 stimulus conditions. For the target and competitor, in Experiment 1, using English sentences, the competitor shared minimally the onset and vowel of the first syllable with the target noun. In Experiment 2, using Mandarin sentences, the competitor shared the pronunciation of the first character (syllable) of the target noun. The distractor was unrelated to both the target and the competitor, neither semantically nor phonetically.

In foil trials, the target words from the critical trials became the competitors, and the competitors from the critical trials became the target words. The distractors in a target trial could also become targets in foil trials, and targets could become distractors in foil trials. Therefore, some foils contain two pairs of targets and competitors (from critical trials), while others have four unrelated words or a pair of competitor and target with two unrelated distractors. Each word appears an equal number of times as a target, foil, or distractor. An example English foil trial consists of words such as *keyring* (target), *slinky*, *liver*, and *litter*. An example Mandarin foil trial consists of words such as *kōng jiě "stewardess"* (target), *lán qiú "basketball"*, *wū méi "plum"*, and *wū guī "tortoise"*.

6.2.2.1 Experiment 3: English stimuli

All sentences and instructions were given in English. An example of a critical trial consists of a target *cabbage*, a competitor *captain*, and two distractors *keyboard*, *corkscrew*. Pictures on screen presented to listeners are given in Figure 6.1.

Figure 6.1: Experiment 3. English stimuli, with pictures presented on the screen to listeners: a target (e.g. *cabbage*), a competitor (e.g. *captain*), and two distractors (e.g. *keyboard* and *corkscrew*).

Based on the manipulation of three binary variables, the example sentences that vary across the two conditions within each variable are listed below. The target word is underlined, and the target word bearing a contrastive pitch accent is capitalized.

information structure

In new information sentences, the competitor was introduced in the first sentence, and the target was introduced in the second sentence as newly introduced information (e.g. *Here is a captain. Susan is going to shred the cabbage after school*). In old information sentences, the target in the second sentence was old information mentioned in the first sentence already (e.g. *Here is a cabbage. Susan is going to shred the cabbage after school*).

Verb appropriateness

In appropriate sentences, the verb in the second sentence was appropriately applied to the target noun as a direct object (e.g. *Here is a captain. Susan is going to shred the cabbage after school*). In inappropriate sentences, the verb was not plausibly applied to the target noun (e.g. *Here is a captain. Susan is going to drink the cabbage after school*).

Pitch accent

In contrastive pitch-accented sentences, a contrastive pitch accent was placed on the target noun (e.g. *Susan is going to shred the CABBAGE after school*). In deaccented sentences, the target carried a neutral accent, and the prosody was a more neutral topic-content melody, with any pitch accent placed on the final adverbial (e.g. *"Susan is going to shred the cabbage after school"*).

6.2.2.2 Experiment 4: Mandarin stimuli

All sentences and instructions were given in Mandarin. An example of a critical trial consists of a target *kōng tiáo "air conditioner"*, a competitor *kōng jiě "stewardess"*, and two distractors *wū guī "tortoise"*, *lán zǐ "basket"*.

Figure 6.2: Experiment 4. Mandarin stimuli, with pictures presented on the screen to listeners: a target (e.g. kōng tiáo "air conditioner"), a competitor (e.g. kōng jiě "stewardess"), and two distractors (e.g. wū guī "tortoise" and lán zǐ "basket").

In the same format as the English stimuli, the following list presents example sentences that vary between the two conditions within each of the three binary variables. The target word is indicated by underlining, and the target word bearing a contrastive pitch accent is in bold.

information structure

In new information sentences, the competitor is named in the first sentence and the target is named in the second sentence as new information:

(1) nǐ kě yǐ kàn dào yī xiē kōng jiě. xiàn zài xiǎo yǔ yào qù ān zhuāng kōng tiáo zhè yàng xià tiān jiù biàn dé liáng kuài le.

You can see some stewardesses. Now Xiaoyu is going to install the air conditioner so that the summer will become cool.

In old/repeated information sentences, the target introduced in the second sentence has already been mentioned in the first sentence:

(2) nǐ kě yǐ kàn dào yī gè kōng tiáo. xiàn zài xiǎo yǔ yào qù ān zhuāng kōng tiáo zhè yàng xià tiān jiù biàn dé liáng kuài le.

You can see an air conditioner. Now Xiaoyu is going to install the air conditioner so that the summer will become cool.

Verb appropriateness

In verb appropriate sentences, the verb in the second sentence is appropriate used with the target noun:

(3) nǐ kě yǐ kàn dào yī gè kōng tiáo. xiàn zài xiǎo yǔ yào qù ān zhuāng kōng tiáo zhè yàng xià tiān jiù biàn dé liáng kuài le.

You can see an air conditioner. Now Xiaoyu is going to install the air conditioner so that the summer will become cool.

In verb inappropriate sentences, the verb in the second sentence is inappropriate used with the target noun:

(4) nǐ kě yǐ kàn dào yī gè kōng tiáo. xiàn zài xiǎo yǔ yào qù wèn le kōng tiáo zhè yàng xià tiān jiù biàn dé liáng kuài le.

You can see an air conditioner. Now Xiaoyu is going to ask the air conditioner so that the summer will become cool.

Pitch accent

In contrastive pitch-accented sentences, the target noun carries contrastive pitch accent:

(5) nǐ kě yǐ kàn dào yī xiē kōng jiě. xiàn zài xiǎo yǔ yào qù ān zhuāng **kōng tiáo** zhè yàng xià tiān jiù biàn dé liáng kuài le.

*You can see some stewardesses. Now Xiaoyu is going to install the **air conditioner** so that the summer will become cool.*

In neutral-accented sentences, the target carried a neutral accent:

(6) nǐ kě yǐ kàn dào yī xiē kōng jiě. xiàn zài xiǎo yǔ yào qù ān zhuāng kōng tiáo zhè yàng xià tiān jiù biàn dé liáng kuài le.

You can see some stewardesses. Now Xiaoyu is going to install the air conditioner so that the summer will become cool.

6.2.3 Procedure

Participants were seated in front of the eye-tracker while wearing headphones. They were informed that, *"we are training a new artificial intelligence to generate sentences. There is a verb and nouns. We want to figure out how easy these sentences can be understood. Your task is to click on the noun that is mentioned in the second phrase. The speed you identify these nouns will let us to measure the comprehensive ability of these artificial generated sentences"* (The data collection was finished before the release of ChatGPT, it was still plausible for AI sentences to sound unnatural). For Mandarin-speaking participants, these instructions were provided in Mandarin to ensure full comprehension.

Data were collected with an Eyelink 1000+ eye-tracker running Experiment Builder software. Gaze data was sampled at 1000Hz. Each experiment began with recalibration, two practice trials, and then the system was recalibrated before the main experiment began. Each participant listened to a total of eighty trials, which were divided into 8 blocks, with each block containing 10 trials. After each block, a short rest period followed and calibration. In this short period of rest, participants were asked to press the keys "1, 2, 3, 4" on the keyboard to indicate how sensible the last trial they had heard was, with 1 denoting the most sensible and 4 indicating the least sensible. This phase was included to assess participants' attention to the trials. Completing this task took each participant approximately thirty minutes.

6.2.3.1 GAMMs analysis

Looks to target and competitor were binned and converted to proportions over 50ms windows, over an interest period stretching from 200 ms before the target onset to 1800ms after target onset. By including a window that starts 200ms before the target onset, it adequately captures anticipatory eye movements that might occur based on the verb semantics preceding the target. Extending the time window to 1800ms after the target onset allows time to observe how eye movements in response to the target word. This duration captures the full trajectory of gaze as listeners identify the target.

In addition to three independent variables, *target advantage* (referred as *TA* in the exper-

iment) was calculated as the dependent variable. In this experiment, TA was calculated by subtracting the proportion of looks to the competitor from proportion of looks to the target. TA was analysed with generalised additive mixed models (GAMMs), using the `mgcv` package (version 1.8.4 (Wood, 2003)) in R (version 4.2.1; (RCoreTeam, 2022)).

Repeated and new information sentences were analysed separately. For both repeated and new sentences, a simple GAMM was built, containing parametric effects and difference smooths for each of the three binary main variables (L1, Accent, Semantics). All factors were treatment-coded, with default levels set in English stimuli as Neutral Accent, Appropriate Verb, and Mandarin L1, and default levels set in Mandarin stimuli as Neutral Accent, Appropriate Verb, and English L1. Interaction terms were added progressively as two-way and then three-way interactions. Each subsequent model was tested against the simpler model with a log-likelihood ratio test, as implemented in the package `itsadug` (version 2.4.1 (van Rij et al., 2022)).

In GAMMs, we care about both parametric terms and smooth terms in understanding the relationships between variables (Sóskuthy, 2017). Parametric terms describe the linear relationships between independent variables and the dependent variable, and reveal the main effects and interactions among the variables. The *estimation* indicates the magnitude and direction (positive/negative) of these effects. The *intercept* represents the baseline value of the dependent variable when all independent variable are at their baseline (reference) levels. Smooth terms, for non-parametric terms, describe the non-linear effects and curvature of the relationships they model. When a smooth term is statistically significant, it indicates the presence of a non-linear effect, and the curvature of the smooth term is significantly different from a flat line (non-linear effect).

Chapter 7

Eye-tracking experiment results and discussion

7.1 Experiment 3: English stimuli

7.1.1 Results

7.1.1.1 Old information

Figure 7.1 shows the gaze traces for old target nouns, while Table 7.1 summarises the final models for old information conditions: the optimal model contained two-way interactions in the parametric terms, and a three-way interaction in the difference smooths.

When processing sentences with old information, Mandarin listeners displayed a significant effect of Accent ($\beta = 0.023, p < .005$). When there was an appropriate verb used preceding the target word, the application of the contrastive pitch accent facilitate the looks to the target. Mandarin listeners also exhibit an effect of Semantics: a reduction in TA for inappropriate verbs ($\beta = -0.060, p < .001$) when a neutral accent was applied on target word. It should be noted that while Mandarin listeners demonstrated sensitivity to both verb appropriateness and pitch accent, the effect of Semantics ($\beta = -0.060, p < .001$) was substantially three times more extreme than that of the Accent ($\beta = 0.023, p < .005$).

Additionally, Mandarin listeners demonstrated an interaction between verb semantics and contrastive pitch accent ($\beta = -0.025, p < .05$). This interaction between contrastive

pitch accent and appropriateness indicated that the use of a contrastive pitch accent exaggerated the inhibitory effect of inappropriate verbs (This is henceforth referred to as the "Semantic deepening effect"). This implies that Mandarin listeners faced greater difficulty processing an inappropriate verb under a contrastive pitch accent compared to a neutral accent.

English listeners demonstrated a higher TA than Mandarin listeners when appropriate verbs were used with a neutral accent ($\beta = 0.265, p < .001$). The lack of interaction between the L1 and accent in GAMMs suggested that, similar to Mandarin listeners, English listeners also show substantial facilitative effect of Accent. This suggests that English listeners experience a substantial increase in TA due to the presence of a contrastive pitch accent on the target word when preceded by an appropriate verb. L1 showed a positive interaction with Semantics ($\beta = 0.032, p < .005$). This means that although English listeners, similar as Mandarin listeners, experienced a reduction in TA to the target when inappropriate verbs were used before the target, this inhibitory effect was less pronounced among English listeners compared to Mandarin listeners ($\beta = -0.060, p < .001$), showing approximately half the amount of inhibition encountered by Mandarin listeners.

Table 7.1: Experiment 3. GAMMs for English sentences with old information. Levels for Accent (abbreviated *Acc*) are neutral (abbreviated *neu*) and target (abbreviated *tar*), with the neutral accent as the default level. Levels for Semantics (*Sem*) are appropriate (reference, *a*) and inappropriate (*i*), with the appropriate semantics being the default level. Levels for L1 are Mandarin (reference, *man*) and English (*eng*), with Mandarin as the default level. Difference smooths are calculated on an 8-level factor representing the interaction between Acc, Sem, and L1.

Old information

Parametric	Est.	SE	t	p
Intercept	0.215	0.031	6.84	$< .001$
Acc=tar	0.023	0.007	3.28	$< .005$
Sem=i	-0.060	0.009	-7.02	$< .001$
L1=eng	0.265	0.042	6.24	$< .001$
Acc=tar:Sem=i	-0.025	0.010	-2.47	$< .05$
Sem=i:L1=eng	0.032	0.010	3.18	$< .005$

Difference smooths	edf	F	p
Window	6.16	10.61	$< .001$
Win:Acc=tar,Sem=a,L1=eng	5.35	8.73	$< .001$
Win:Acc=neu,Sem=i,L1=eng	4.16	3.53	$< .005$
Win:Acc=tar,Sem=i,L1=eng	6.20	7.65	$< .001$
Win:Acc=neu,Sem=a,L1=man	1.01	0.43	.52
Win:Acc=tar,Sem=a,L1=man	1.00	6.04	$< .05$
Win:Acc=neu,Sem=i,L1=man	2.66	3.58	$< .01$
Win:Acc=tar,Sem=i,L1=man	2.61	5.89	$< .001$
Window, by subj	489.18	8.97	$< .001$

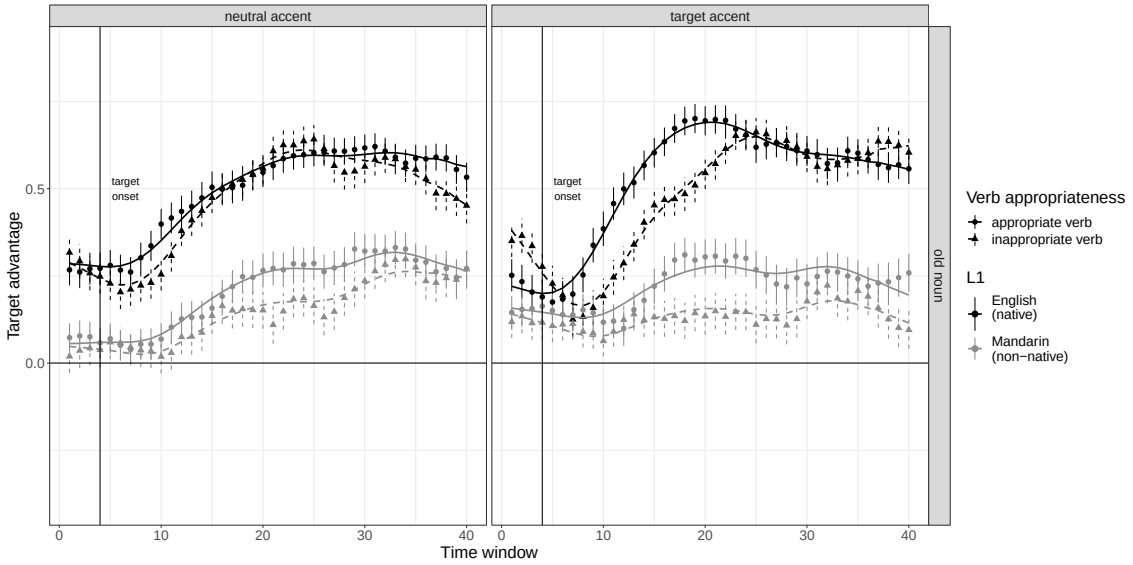


Figure 7.1: Experiment 3. English stimuli, gaze traces with old nouns, with English listeners as native listeners and Mandarin listeners as non-native listeners. *Target Advantage* (TA) was compared across English (black line) and Mandarin (grey line) listeners, for appropriate (dotted circles, connected by solid lines) and inappropriate (solid triangle, connected by dashed lines) verbs semantics. *Neutral accent* is on the left, while *Target accent* (contrastive pitch accent) is on the right. Target accent is the unexpected prosody for target noun as old information.

The absence of a three-way interaction suggests that, similar to Mandarin listeners, English listeners also exhibited an exaggerated negative and inhibitory effect of inappropriate verb semantics under a contrastive pitch accent compared to a neutral accent. The significantly different curvature of the gaze traces indicates that this appropriateness effect was concentrated within the first 25 time windows for English listeners (Figure 7.1, right black lines), whereas for Mandarin listeners, this effect could extend beyond 30 time windows.

The shape of the curve is also noteworthy. As demonstrated by Figure 7.1, before target onset, all TA values are positive, suggesting a tendency to anticipate a repetition of old information for both native and non-native listeners. Under a neutral accent, which is consistent with old information, there is not much decrease in TA even after inappropriate verbs. However, under a contrastive pitch accent, there is a distinct drop after inappropriate verbs for native listeners, illustrating the moment of deepened semantic processing.

In other words, when processing sentences with repeated information, both Mandarin and English listeners exhibited a similar pattern in their utilization of cues. They were both sensitive to the contrastive pitch accent applied to the target noun as well as to the semantics of the verb preceding the target noun. However, Mandarin listeners experienced a more pronounced negative effect from the use of inappropriate verbs compared to English listeners. Moreover, both groups demonstrated an interaction between the contrastive pitch accent and verb semantics, as they both encountered an intensified negative effect from inappropriateness when the contrastive pitch accent was applied to the target.

7.1.1.2 New information

Figure 7.2 shows the gaze traces for new target nouns, while Table 7.2 summarizes the final models for new information conditions: the model ended up needing three-way interactions in both parametric terms and smooths.

When processing new information sentences, Mandarin listeners also demonstrated a response pattern that was clearly slower than that of English listeners. English listen-

ers reached peak TA at about Time Window 25, while Mandarin listeners did not peak until Time Window 35. For Mandarin listeners, there was a positive effect of Accent ($\beta = 0.044, p < .001$), which means Mandarin listeners benefited from the addition of a contrastive pitch accent on the target word when processing new information sentences with an appropriate verb.

Nevertheless, there was little effect of Semantics observed ($\beta = 0.001, p = .89$). This suggests that the use of an inappropriate verb with a neutrally accented target did not substantially reduce the TA for Mandarin listeners. The gaze traces showed distinct curvatures: the left grey lines (Figure 7.2), indicating the appropriate verb condition (dot line) rose faster and peaked higher, but started with a slightly lower TA than the inappropriate verb condition (triangle line). This rise canceled out the initial disadvantage, resulting in a null parametric effect of Semantics.

Further, Mandarin listeners exhibited an interaction between verb semantics and contrastive pitch accent. With the presence of pitch accent, the impact of an inappropriate verb (Semantic effect) became more substantial ($\beta = -0.030, p < .05$).

Table 7.2: Experiment 3. GAMMs for English sentences with new information. Levels for Accent (abbreviated *Acc*) are neutral (abbreviated *neu*) and target (abbreviated *tar*), with the neutral accent as the default level. Levels for Semantics (*Sem*) are appropriate (reference, *a*) and inappropriate (*i*), with the appropriate semantics being the default level. Levels for L1 are Mandarin (reference, *man*) and English (*eng*), with Mandarin as the default level. Difference smooths are calculated on an 8-level factor representing the interaction between Acc, Sem, and L1.

New information

Parametric	Est.	SE	t	p
Intercept	0.143	0.029	5.00	$< .001$
Acc=tar	0.044	0.010	4.61	$< .001$
Sem=i	0.001	0.010	0.13	.89
L1=eng	0.263	0.040	6.54	$< .001$
Acc=tar:Sem=i	-0.030	0.014	-2.14	$< .05$
Acc=tar:L1=eng	-0.069	0.013	-5.12	$< .001$
Sem=i:L1=eng	-0.071	0.013	-5.29	$< .001$
Acc=tar:Sem=i:L1=eng	0.043	0.019	2.21	$< .05$
Difference smooths		edf	F	p
Window		8.17	83.30	$< .001$
Win:Acc=tar,Sem=a,L1=eng		1.00	5.23	$< .05$
Win:Acc=neu,Sem=i,L1=eng		8.12	40.46	$< .001$
Win:Acc=tar,Sem=i,L1=eng		7.97	38.80	$< .001$
Win:Acc=neu,Sem=a,L1=man		5.67	23.83	$< .001$
Win:Acc=tar,Sem=a,L1=man		5.75	21.84	$< .001$
Win:Acc=neu,Sem=i,L1=man		5.58	25.05	$< .001$
Win:Acc=tar,Sem=i,L1=man		5.75	26.21	$< .001$
Window, by subj		498.57	8.16	$< .001$

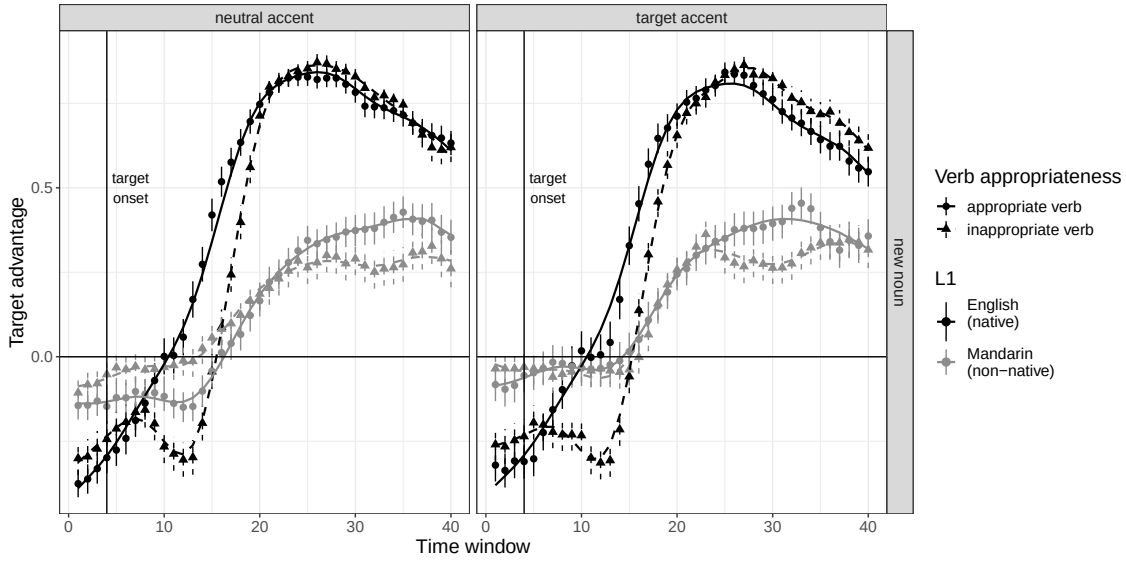


Figure 7.2: Experiment 3. English stimuli, gaze traces with new nouns, with English listeners as native listeners and Mandarin listeners as non-native listeners. *Target Advantage* (TA) was compared across English (black line) and Mandarin (grey line) listeners, for appropriate (dotted circles, connected by solid lines) and inappropriate (solid triangle, connected by dashed lines) verbs semantics. *Neutral accent* is on the left, while *Target accent* (contrastive pitch accent) is on the right. Target is the expected prosody for target noun as new information.

With target noun as new information, English listeners exhibited a higher TA than Mandarin listeners when appropriate verbs were presented with a neutral accent ($\beta = 0.263, p < .001$). English listeners showed a negative interaction term for the effect of Accent on appropriate verbs ($\beta = -0.069, p < .001$), which was more pronounced than the facilitation of the contrastive pitch accent observed in Mandarin listeners. Thus, in contrast to Mandarin listeners, English listeners did not benefit from the contrastive pitch accent on the target, even though it correctly signaled the new information structure. In fact, their TA has even been reduced.

L1 also exhibited a significant interaction with Semantics when presented with a neutral accent for English listeners ($\beta = -0.071, p < .001$). This means that inappropriate verb semantics reduced TA more for English listeners than for Mandarin listeners.

Further, for English listeners, there was a three-way interaction between L1, Accent, and Semantics ($\beta = 0.043, p < .05$). The positive value of the coefficient means that the negative effect of Semantics observed with a neutral accent for English listeners was lessened by the presence of a contrastive pitch accent on the target. However, this cannot merely mean that the contrastive pitch accent facilitated the identification of targets after inappropriate verbs for English listeners. There is a nuanced interpretation, that is, pitch accent alters the timing and possibly the depth of processing that listeners engage in when encountering semantically challenging sentences. As shown by Figure 7.2, there is a slight delayed peak in TA for an inappropriate verb under pitch accent, suggesting it causes listeners to spend more time processing the target noun, leading to a slower identification of the target noun. After this delayed peak, the TA for the inappropriate verb with target accent surpasses that of the appropriate verb without pitch accent. This indicates a compensatory effect, where initial confusion or increased processing time due to the accent and semantic mismatch eventually leads to better comprehension and slower looking away. The pitch accent may initially increase cognitive load, leading to slower processing, but it also keeps the listener engaged with the target noun longer, potentially leading to more thorough processing or a stronger eventual recognition of the target. The explanation implies that the pitch accent does more than simply aid in understanding;

rather, it could modify how semantics are processed depending on the prosodic context.

Before the target onset, both groups of listeners started below 0, indicating an expectation of old information. Since the competitor represents the old information, this led to more looks at the competitor than at the target before the target onset, resulting in a negative TA. For native listeners, under the inappropriate verb condition, there was an immediate drop in TA right after the target onset. This indicates a clear effect of Semantics. However, this pattern was not observed in non-native listeners. The effect of Semantics only emerged quite late, suggesting delayed semantic processing in non-native listeners.

In other words, when processing sentences with new information, Mandarin listeners showed insensitivity to verb semantics but demonstrated sensitivity to prosody, which facilitated TA more than observed in English listeners. However, the effect of semantics was modulated by the contrastive pitch accent; Mandarin listeners exhibited a deeper effect of Semantics when the contrastive pitch accent was applied to the target. In contrast, English listeners displayed sensitivity to both verb semantics and prosodic accent: applying either an inappropriate verb or pitch accent to the target noun resulted in reduced TA for English listeners. Moreover, these Semantic and Accent effects were more negatively pronounced for English listeners than for Mandarin listeners. Furthermore, English listeners exhibited an interaction between verb semantics and contrastive pitch accent. Although this interaction was positive, as showed in Figure 7.2, applying an inappropriate verb with a pitch accent resulted in slower detection of the target noun and slower looking away compared to the neutral accent condition.

7.1.2 Discussion

With English sentences as stimuli completed by native English listeners and native Mandarin listeners learning English as L2, Experiment 3 confirms some, but not all, of the predictions. Generally, Mandarin and English listeners employed similar speech perceptual strategies when processing sentences containing old information. However, they adopted markedly different strategies when processing sentences containing new information.

7.1.2.1 Old information

The results for old information confirm the prediction that Mandarin listeners, as non-native listeners in the context of English, attended to both semantic information (Semantic effect) and prosodic information (Accent effect) when processing sentences with repeated information as the target object. However, they exhibited greater sensitivity to semantic information than to prosodic information, as showed by the larger Semantic effect compared to the Accent effect. This aligns with the prediction for processing sentences with repeated information.

However, it was not predicted that Mandarin listeners would be capable of integrating both semantic and prosodic information when processing sentences with old information. Moreover, Mandarin listeners adopted similar strategies for cue utilization and integration as English listeners: when both groups listened to sentences with repeated information, adding a contrastive pitch accent increased attention to semantics. This was predicted for English listeners but not for Mandarin listeners. This replicates a pattern observed in a previous studies (Brunellière et al., 2019; Li & Lu, 2011; Li & Ren, 2012; L. Wang et al., 2011), where placing a contrastive pitch accent on a target noun deepened semantic processing.

Therefore, some predictions were confirmed while others were not. For sentences containing old information, both Mandarin and English listeners adopted a similar speech perceptual strategy: attending to both verb semantics and prosodic information and flexibly integrating these cues in a comparable manner. The pitch accent on the target noun facilitated target detection, while the inappropriate verb made it more difficult to detect the target noun, with English listeners being more affected than Mandarin listeners by the inappropriate verb that preceded the target noun. Furthermore, the presence of a pitch accent exacerbated the effect of the inappropriate verb, suggesting that the effectiveness of one cue is influenced by the presence of the other for both groups of listeners.

7.1.2.2 New information

When processing sentences with new information, it was predicted that English listeners would also attend to both semantic and prosodic information and integrate these cues. However, there were results that were not predicted, and Mandarin and English listeners adopted very different speech perceptual strategies for processing new information sentences, as sentences containing new information posed a greater challenge in target detection compared to sentences with old information.

In the new information sentence, although English listeners utilized both types of cues and showed integration, it was unexpected that the utilization and integration of these cues were different from how they processed old information sentences. It was surprising to find that English listeners did not benefit from the contrastive pitch accent to detect the target as Mandarin listeners, although the accent had the correct positioning to indicate the information structure. Perhaps this occurred because the prosodic pattern, by virtue of its unexpectedness, served to direct their attention to the target noun. In new information sentences, the prosodic pattern was expected, and as a result, almost invisible to English listeners. Further, the integration of semantics and prosody differs in sentences containing old information: the negative effect of Semantics was mitigated by the presence of a contrastive pitch accent, as observed by Ventura et al (2020), in which the post-focal position attenuated N400 congruence effects, suggesting that the incongruence in the post-focal position may still undergo deep semantic processing, albeit to a lesser extent than in the focal position. Therefore, it is reasonable to infer that the negative effect of an inappropriate verb in the "pre-focus" position was diminished and attenuated by the prosodic prominence at the target noun, resulting in the reduced negative effect of Semantics when paired with a contrastive pitch accent on the target.

For Mandarin listeners as non-native listeners, findings were also not predicted in the new information sentences: Mandarin listeners did prioritize semantic information over prosodic information when processing target words as old information but reversed this prioritization, giving preference to prosody over semantics, when processing target words as new information. The pitch accent aided perception generally in new information

sentences, where that prosodic structure was appropriate to the information structure of the sentence. This is perhaps because the accent rendered the target noun louder and more salient to detect. Further, it was also not predicted that even processing new information sentences, Mandarin listeners were still able to integrate semantics and prosody: the effect of Semantics became more substantial when it interacted with the contrastive pitch accent, suggesting the utilization of either the semantics or prosody is affected by one another for Mandarin listeners.

Therefore, Mandarin and English listeners adopted very different speech perceptual strategies for processing new information sentences. When processing new information sentences, Mandarin listeners focused their attention solely on a single cue: prosodic accent, while relying little on verb semantics. English listeners demonstrated their ability to allocate attention weight to both verb semantics and prosodic accent. Both groups show a pattern where one cue is affected by another, showing integration of cues. However, the ways of interaction were different for Mandarin and English listeners. For Mandarin listeners, the three-way interaction term was negative, suggesting that the negative effect of Semantics was little, but became more substantial under the impact of the pitch accent, although it did not substantially reduce TA. For English listeners, the pitch accent further slowed down the detection of target noun with inappropriate verb, as showed in the delayed peak in the Figure 7.2, perhaps the pitch accent attracts attention to the verb-target semantic mismatch. Therefore, the strategy to integrating multiple cues when processing new information sentences differed between English and Mandarin listeners.

7.2 Experiment 4: Mandarin stimuli

7.2.1 Results

7.2.1.1 Old information

Figure 7.3 shows the gaze traces for old target nouns, while Table 7.3 summarises the final models for old information conditions: the optimal model contained two-way interactions in the parametric terms, and a three-way interaction in the difference smooths.

In the case of old information sentences, English listeners exhibited a positive effect of Accent ($\beta = 0.038, p < .001$), suggesting that the application of a contrastive pitch accent on the target noun enhanced TA for English listeners. Additionally, English listeners experienced a significant negative effect of Semantics ($\beta = -0.059, p < .001$), whereby the use of inappropriate verb semantics before the target noun reduced TA.

Table 7.3: Experiment 4. GAMMs for Mandarin sentences with old information. Levels for Accent (abbreviated *Acc*) are neutral (abbreviated *neu*) and target (abbreviated *tar*), with the neutral accent as the default level. Levels for Semantics (*Sem*) are appropriate (reference, *a*) and inappropriate (*i*), with the appropriate semantics being the default level. Levels for L1 are Mandarin (reference, *man*) and English (*eng*), with English as the default level. Difference smooths are calculated on an 8-level factor representing the interaction between Acc, Sem, and L1.

Old information

Parametric	Est.	SE	t	p
Intercept	0.228	0.028	8.17	$< .001$
Acc=tar	0.038	0.009	4.14	$< .001$
App=i	-0.059	0.009	-6.53	$< .001$
L1=man	0.106	0.035	3.04	$< .005$
Acc=tar:App=i	-0.079	0.013	-6.17	$< .001$

Difference smooths	edf	F	p
Window	4.41	8.56	$< .001$
Win:Acc=tar,Sem=a,L1=eng	1.00	0.37	.54
Win:Acc=neu,Sem=i,L1=eng	1.01	3.27	.07
Win:Acc=tar,Sem=i,L1=eng	3.77	4.31	$< .001$
Win:Acc=neu,Sem=a,L1=man	3.62	5.78	$< .001$
Win:Acc=tar,Sem=a,L1=man	1.96	1.80	.12
Win:Acc=neu,Sem=i,L1=man	5.63	9.97	$< .001$
Win:Acc=tar,Sem=i,L1=man	4.22	6.94	$< .001$
Window, by subj	302.83	5.77	$< .001$

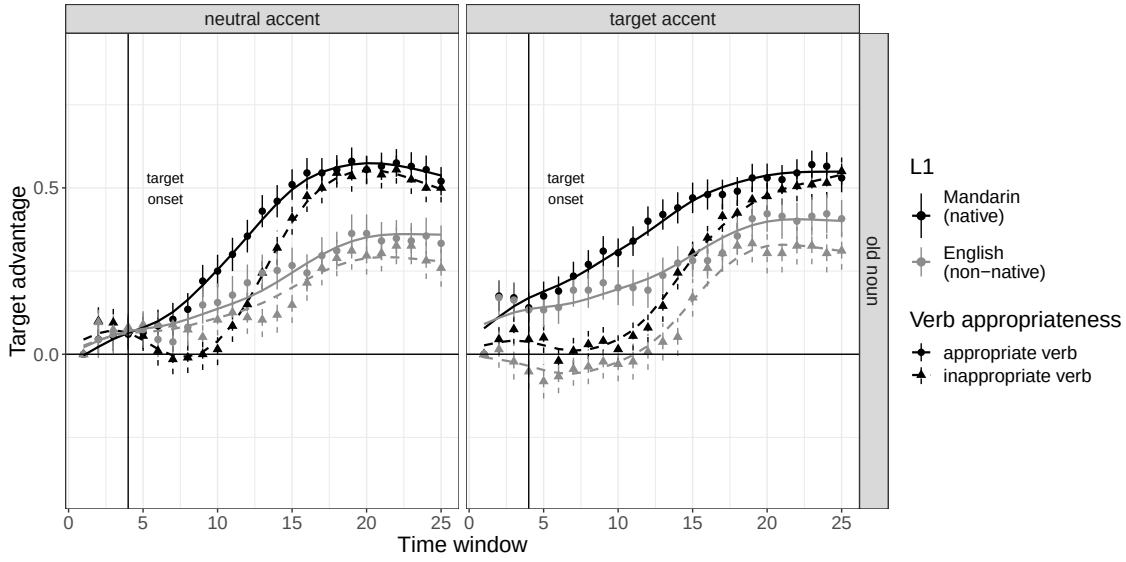


Figure 7.3: Experiment 4. Mandarin stimuli, gaze traces with old nouns, with Mandarin listeners as native listeners and English listeners as non-native listeners. *Target Advantage* (TA) was compared across Mandarin (black line) and English (grey line) listeners, for appropriate (dotted circles, connected by solid lines) and inappropriate (solid triangle, connected by dashed lines) verbs semantics. *Neutral accent* is on the left, while *Target accent* (contrastive pitch accent) is on the right. Target accent is the unexpected prosody for target noun as old information.

A negative interaction term between Accent and Semantics was observed for English listeners ($\beta = -0.079, p < .001$). This interaction indicates that contrastive pitch accent on target noun even exaggerated the negative Semantics effect for English listeners.

When compared to English listeners, Mandarin listeners consistently demonstrated a higher overall TA across the conditions for sentences containing old information ($\beta = 0.106, p < .005$). No interactions emerged between L1 and any of the other effects, indicating that Accent, Semantics, and the Semantic deepening effect applied similarly to Mandarin and English listeners.

As showed in Figure 7.3, the curve shapes are also worth noting. For Mandarin listeners (black lines), there is a quick dip in TA immediately after target onset, indicating that Mandarin listeners responded very quickly to the mismatch between verb semantics and the target noun. This quick response is absent in the grey curves for English listeners, suggesting that English listeners as non-native listeners were not as quick to notice the mismatch, which is consistent with slower semantic processing for non-native listeners. Additionally, the lines for appropriate and inappropriate verb semantics converge sooner for Mandarin than for English listeners—by window 18 in the neutral accent condition, and not much later under the contrastive pitch accent. In contrast, English listeners' lines remain well separated through time window 25, providing further evidence of slower semantic processing for English listeners.

In other words, when dealing with sentences containing old information, both English and Mandarin listeners demonstrated effect of Accent and Semantics, with positive impact of pitch accent in increasing TA and negative effect of inappropriate verb in reducing TA. Further, both groups exhibited an interaction between Accent and Semantics, with the contrastive pitch accent on the target noun amplifying the negative effect of Semantics, thereby further reducing TA.

7.2.1.2 New information

Figure 7.4 shows the gaze traces for new target nouns, while Table 7.4 summarizes the final models for new information conditions: the model ended up needing three-way interactions in both parametric terms and smooths.

When dealing with new information sentences, English listeners exhibited an effect of Accent ($\beta = -0.036, p < .05$); however, it was negative. This negative effect indicated that, despite the contrastive pitch accent correctly signaling the new information structure of the target noun, English listeners did not benefit from the presence of the accented target noun but were adversely affected. This result differs from previous studies on the "search for focus effect" (Akker & Cutler, 2003; Ip & Cutler, 2020), where listeners benefited from the contrastive pitch accent in detecting the target noun as new information.

English listeners also showed an effect of Semantics ($\beta = -0.15, p < 0.001$). This suggests that in the presence of a neutral accent, the inappropriate verb also led to a reduction in TA. In comparison to the effect of Accent ($\beta = -0.036, p < .05$), the effect of Semantics ($\beta = -0.15, p < 0.001$) observed in English listeners was substantially larger. This suggests that although English listeners were sensitive to both semantic and prosodic cues, they could demonstrate more sensitivity to Semantics than to Accent.

Further, English listeners did not show an interaction between Accent and Semantics ($\beta = 0.039, p = .07$), indicating that verb semantics and prosodic accent did not mutually influence each other. This means there was no Semantic deepening effect in this new information condition for English listeners.

Table 7.4: Experiment 4. GAMMs for Mandarin sentences with new information. Levels for Accent (abbreviated *Acc*) are neutral (abbreviated *neu*) and target (abbreviated *tar*), with the neutral accent as the default level. Levels for Semantics (*Sem*) are appropriate (reference, *a*) and inappropriate (*i*), with the appropriate semantics being the default level. Levels for L1 are Mandarin (reference, *man*) and English (*eng*), with English as the default level. Difference smooths are calculated on an 8-level factor representing the interaction between Acc, Sem, and L1.

New information

Parametric	Est.	SE	t	p
Intercept	0.202	0.030	6.67	$< .001$
Acc=tar	-0.036	0.015	-2.39	$< .05$
Sem=i	-0.15	0.015	-9.97	$< .001$
L1=man	0.079	0.038	2.06	$< .05$
Acc=tar:Sem=i	0.039	0.021	1.84	.07
Acc=tar:L1=man	0.089	0.019	4.60	$< .001$
Sem=i:L1=man	-0.013	0.019	-0.65	.51
Acc=tar:Sem=i:L1=man	-0.108	0.027	-3.92	$< .001$
Difference smooths		edf	F	p
Window		6.86	16.93	$< .001$
Win:Acc=tar,Sem=a,L1=eng		1.29	1.79	.10
Win:Acc=neu,Sem=i,L1=eng		3.26	6.27	$< .001$
Win:Acc=tar,Sem=i,L1=eng		3.34	7.20	$< .001$
Win:Acc=neu,Sem=a,L1=man		1.00	14.78	$< .001$
Win:Acc=tar,Sem=a,L1=man		1.00	19.19	$< .001$
Win:Acc=neu,Sem=i,L1=man		5.14	32.93	$< .001$
Win:Acc=tar,Sem=i,L1=man		4.75	22.34	$< .001$
Window, by subj		292.12	5.04	$< .001$

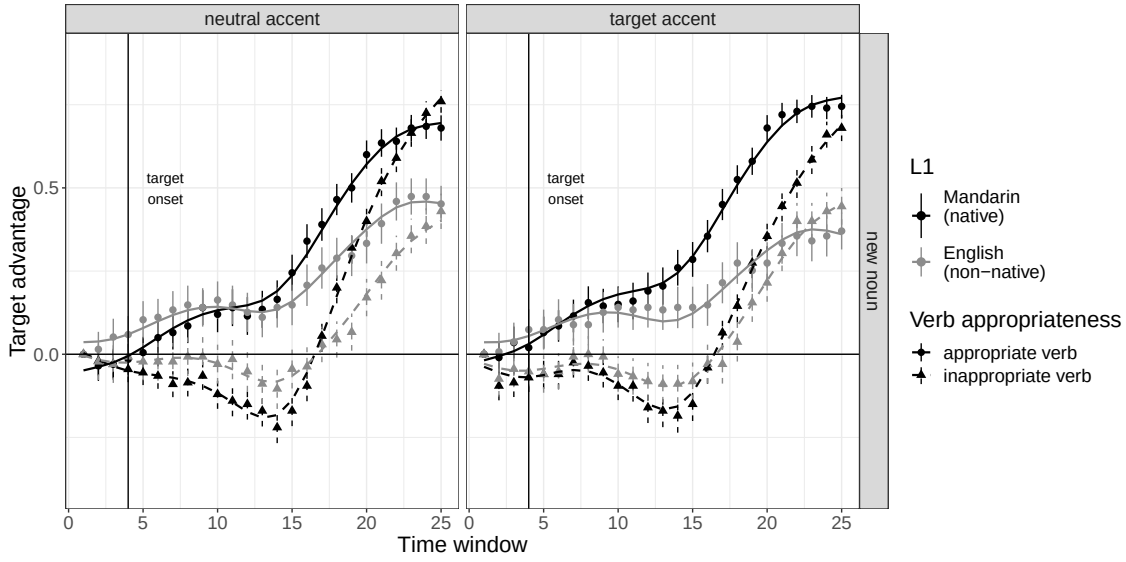


Figure 7.4: Experiment 4. Mandarin stimuli, gaze traces with new nouns, with Mandarin listeners as native listeners and English listeners as non-native listeners. *Target Advantage* (TA) was compared across Mandarin (black line) and English (grey line) listeners, for appropriate (dotted circles, connected by solid lines) and inappropriate (solid triangle, connected by dashed lines) verbs semantics. *Neutral accent* is on the left, while *Target accent* (contrastive pitch accent) is on the right. Target accent is the expected prosody for target noun as new information.

Similar to the results from Experiment 3, Mandarin listeners as native listeners exhibited an overall higher TA than English listeners as non-native listeners when processing new information sentences ($\beta = 0.079, p < .05$).

The positive interaction between L1 and Accent ($\beta = 0.089, p < .001$) was large enough to reverse the negative Accent effect observed in English listeners ($\beta = -0.036, p < .05$). This suggests that Mandarin listeners showed the expected facilitation of contrastive pitch accent as it correctly signaled the new information structure. However, the lack of significant interaction between L1 and Semantics ($\beta = -0.013, p = .51$) reveals that Mandarin and English listeners have similar Semantics effect.

Moreover, a negative three-way interaction emerged between L1, Accent and Semantics ($\beta = -0.108, p < .001$). This suggests that, unlike English listeners, Mandarin listeners did show the Semantic deepening effect. In other words, prosodic accent on the target word did not change the Semantic effect for English listeners, but it exaggerated the negative Semantics effect for Mandarin listeners.

The curves are also interesting to note in Figure 7.4, TA starts at approximately 0 for both Mandarin and English listeners. This is in stark contrast to Experiment 3, where the curves for new information began significantly below 0. This difference suggests that in the Mandarin context, listeners may be less likely to assume continued use of old information. Both groups of listeners displayed a noticeable and sharp increase in their curves around window 14, likely corresponding to the time needed to disambiguate the target noun from its competitor. The flat segment before this rise reflects the role of verb semantics, with listeners maintaining stable choices between the target and competitor until phonetic confirmation from the second syllable of the target noun is received. This confirmation likely occurs at the point of rising, representing the moment of phonetic disambiguation, while the separation between appropriate and inappropriate lines before the rise probably reflects the contribution of semantics alone.

In other words, English listeners remained sensitive to both prosodic accent and verb semantics when processing new information sentences. However, the Semantic deepening

effect disappeared for English listeners, and they did not exhibit an interaction between prosodic accent and verb semantics. This suggests a reduced ability in integrating these cues when encountering new information sentences. On the other hand, Mandarin listeners attended to both prosodic accent and verb semantics when processing new information sentences, and they demonstrated an interaction between prosodic accent and verb semantics. This interaction resulted in a greater reduction in TA compared to English listeners when facing new information sentences containing an inappropriate verb used with an accented target noun. The semantics deepening effect persisted in the case of Mandarin listeners when presented with new information sentences. However, Mandarin listeners showed a larger effect of Accent than English listeners and exhibited a similar magnitude of effect of Semantics to that observed in English listeners.

7.2.2 Discussion

With Mandarin sentences as stimuli completed by native Mandarin listeners and native English listeners learning Mandarin as L2, Experiment 4 confirms some predictions, but with unexpected results. In Mandarin context, English and Mandarin listeners employed a similar speech perceptual strategy when processing old information sentences, but they also adopted notably different strategies when processing new information sentences.

7.2.2.1 Old information

When processing repeated information Mandarin sentences, it was predicted that Mandarin listeners, as native speakers, would be sensitive to both semantic and prosodic information and capable of integrating these cues during real-time speech processing.

However, it was not predicted that English listeners, as non-native speakers in the Mandarin context, would demonstrate a similar perceptual and processing strategy as Mandarin listeners. English listeners benefited from the pitch accent (positive effect of Accent) but were inhibited by the inappropriate verb (negative effect of Semantics). Additionally, they integrated these cues by showing that the negative effect of the inappropriate verb was exacerbated by the contrastive pitch accent when processing old information sentences.

Thus, some predictions were confirmed while others were not. In sentences containing old information, both Mandarin and English listeners employed a comparable approach to speech perception: they were sensitive to and relied on both semantics and prosody, integrating them in the same way. The contrastive pitch accent applied to the target noun facilitated detection, while the use of an inappropriate verb hindered the identification process. Moreover, the impact of the pitch accent worsened the adverse effect of the inappropriate verb when combined, indicating that the utilization of either cue is influenced by the presence of the other for both groups of listeners.

7.2.2.2 New information

When processing new information Mandarin sentences, English and Mandarin listeners adopted divergent strategies. It was predicted that Mandarin listeners as native listeners would consistently show sensitivity to both semantics and prosody, exhibiting a positive effect of contrastive pitch accent and a negative effect of inappropriate verb semantics. Mandarin listeners also integrated these cues during the processing of new information sentences, as evidenced by the exaggeration of the negative effect of inappropriate verb semantics by the contrastive pitch accent applied to the target noun. The processing strategy applied by Mandarin listeners in new information sentences remained consistent with that observed in old information sentences.

For English listeners processing new information sentences, English listeners showed negative effect of Accent: contrastive pitch accent on the target noun inhibited their detection, even though the contrastive pitch accent correctly signaled the information structure of the target noun in the new information condition. This difference is distinct from their processing of old information sentences and also differs from Mandarin listeners as non-native listeners processing of new information sentences in Experiment 3.

However, the results did confirm the prediction that English listeners, as non-native listeners, showed no interaction between Semantics and Accent during the processing of new information sentences, suggesting a reduced ability of non-native listeners to integrate multiple cues compared to native listeners.

Therefore, Mandarin and English listeners adopted very different speech perceptual strategies for processing new information sentences. When processing new information sentences, Mandarin listeners demonstrated an similar perceptual strategy and manner as they did when processing old information sentences. In contrast, English listeners demonstrated sensitivity to both verb semantics and prosodic pitch accent, but with a negative impact received from the contrastive pitch accent applied to the target. Further, Mandarin listeners were capable of integrating verb semantics and prosody together when processing new information sentences, while English listeners were not able to demonstrate this integration, suggesting that English listeners as non-native listeners utilized cues during new information processing independently and were less able to combine them.

7.3 Discussion

The primary goal of Experiment 3 and 4 was to assess the ability of both native and non-native listeners to utilize and integrate multiple cues, shedding light on their perceptual flexibility during real-time sentence processing.

I expected that native listeners would demonstrate higher perceptual flexibility, being sensitive to the cues used in the experiment and adeptly integrating them. In contrast, I also expected that non-native listeners would exhibit lower perceptual flexibility, showing greater sensitivity to specific cues and struggling to integrate multiple cues effectively during real-time speech processing and comprehension.

The overall TA for non-native listeners was lower compared to native listeners. Non-native listeners typically exhibited later peak than native listeners across both Mandarin and English stimulus contexts. However, as discussed in the preceding *Sections 7.1* and *7.2*, divergent perceptual strategies and patterns emerged not only among native and non-native listeners but also under different language stimuli. This reflects that not only the non-native language context but also the distinct linguistic natures of the stimulus languages (English and Mandarin) can influence listeners' speech perceptual strategies.

7.3.1 Perceptual flexibility under old and new information

The manifestation of *high perceptual flexibility* or *reduced perceptual flexibility*, and whether non-native listeners can approach a level of perceptual flexibility close to that of native listeners, depends on whether the sentences contain old or new information.

When processing old information sentences, in both parts of the experiment employing either English or Mandarin stimuli, both native and non-native listeners consistently adopted highly similar perceptual strategies. In other words, they utilized and integrated cues in the same way. This indicates that when processing old information, both native and non-native listeners exhibited a comparable perceptual strategy. This consistency was also observed between the Mandarin and English stimulus contexts.

Both native and non-native listeners showed sensitivity to and reliance on both verb semantics and prosodic pitch accent. They demonstrated the positive effect of a contrastive pitch accent applied to the target noun and the negative effect of inappropriate verb semantics preceding the target noun. Additionally, they exhibited integration of these cues by emphasizing the negative effect of the inappropriate verb for detecting the target noun when a contrastive pitch accent was added to the target noun. However, it is worth noting that non-native listeners consistently showed greater sensitivity to verb semantics than to prosody, as demonstrated by a greater decrease in TA compared to prosody.

On the other hand, when listeners processed sentences containing new information, a noticeable contrast emerged between native and non-native listeners in both sections of the experiment using English and Mandarin stimuli. They exhibited distinct approaches to assigning attention weight and integrating multiple cues during real-time speech perception of new information.

Under English context, when processing new information sentences, contrastive pitch accent on target noun facilitates detection for Mandarin listeners. However, they did not rely too much on verb semantics, possibly because the verb semantics with target as new information created additional processing difficulty for Mandarin listeners. Besides, the

contrastive pitch accent on the target made it easier to detect, especially when encountered as new information. In contrast, English listeners showed both Semantics and Accent effects. The inappropriate verb semantics also exerted a negative effect on English listeners, similar to Mandarin listeners, but with a different impact from the prosodic pitch accent. This suggests that the contrastive pitch accent may interfere with the detection of new information by drawing excessive attention to the target noun and potentially overshadowing or diverting attention away from other cues. This overemphasis may create cognitive interference or disruption.

Further, the ways of integrating cues also differed between English and Mandarin listeners when listening to English sentences. While Mandarin listeners relied on the contrastive pitch accent to facilitate detection, the addition of inappropriate verb semantics reduced this benefit from prosodic information. In contrast, the negative effect of the prosodic pitch accent became positive when combined with inappropriate verb semantics for English listeners. This suggests that when processing inappropriate verb semantics, English listeners started to utilize prosodic information on the target noun to facilitate detection.

Under the Mandarin context, listeners also exhibited differing speech perceptual strategies when processing new information sentences. Mandarin listeners maintained a consistent perceptual strategy similar to that used for processing old information sentences, emphasizing prosodic accent. In contrast, English listeners showed sensitivity to both verb semantics and prosodic pitch accent but struggled with integrating these cues effectively. Additionally, the contrastive pitch accent on the Mandarin target noun inhibited English listeners' detection of the target noun.

Therefore, it is evident that when processing old information sentences, regardless of the language stimulus contexts, non-native listeners were capable of demonstrating a comparable perceptual flexibility to native listeners. They showed sensitivity to the cues applied in the current eye-tracking experiment and integrated multiple cues in the same way as native listeners. However, regardless of language stimulus contexts, non-native listeners also exhibited reduced perceptual flexibility compared to native listeners when processing

new information sentences. This was evidenced by their tendency to abandon certain cues during new information processing. This could be due to limited L2 proficiency and increased cognitive load. For example, in Experiment 3, Mandarin listeners often abandon verb semantic information when processing English new information sentences, and they have a lesser ability to integrate multiple cues compared to native listeners, such as English listeners failing to integrate verb semantics and prosody when processing Mandarin new information sentences in Experiment 4.

7.3.2 Perceptual flexibility under L1- and L2-specific influence

The manifestation of *high perceptual flexibility* or *low perceptual flexibility*, and whether non-native listeners can achieve a similar level of perceptual flexibility as native listeners, depends on both the language being processed (L2 influence) and their native language background (L1 influence). In other words, listeners adopt different perceptual strategies depending on the language they are processing, and these strategies are influenced by their native language. They transfer their native language's perceptual proficiency in certain cues to non-native speech perception (see *cue weighting theory* in Tremblay et al., 2018, also Francis et al., 2008; Iverson et al., 2003; Qin et al., 2017), recall that Dutch listeners used F0 rise earlier and to a greater extent than English listeners. This is attributed to the greater functional weight of the F0 rise in Dutch compared to English (Tremblay et al., 2018). The results of current eye-tracking experiments demonstrate that listeners can also transfer the inability to perceive certain cues in their L2.

The characteristics of the stimulus language can have either a positive or negative impact on the performance of non-native listeners. For example, the divergent prosodic system of English and Mandarin may account for the different utilization of prosodic information for English and Mandarin listeners under the English and Mandarin stimulus contexts. Under the English context, when Mandarin listeners processed English new information sentences, non-native Mandarin listeners processing English sentences still derived benefits from contrastive focus prosody even in new information context, while under the Mandarin context, when English listeners processed Mandarin stimuli for new information, they showed that they were negatively affected by the contrastive pitch accent applied on

the target noun, although the contrastive pitch accent accurately signaled the information structure in this situation. This divergence may reflect how pitch accent operates in Mandarin compared to English and the differences between Mandarin and English listeners' use of prosodic accent for contrastive focus.

As outlined in Section 5.4 *Utilization and integration of pitch accent in relation with information structure*, Mandarin and English exhibit different pitch patterns. Mandarin uses a contour tone system where pitch patterns shift between levels, contrasting with English's use of rising and falling intonation patterns to signal contrastive pitch accent. Mandarin's contrastive pitch accent operates at the lexical level, altering pitch contours and thus word meanings significantly. In contrast, English emphasizes prosody at the sentence or phrase level, with pitch accents primarily highlighting significance, conveying mood, signaling questions, or emphasizing contrast without fundamentally altering word meanings. Therefore, Mandarin's contrastive pitch accent system more directly impacts word meanings due to its tone system (Shih, 2004): lexical tones interact with intonation when the sentence contains the contrastive focus. The tones before the narrow focus become weak, and their effects are not strong enough to have an impact beyond the narrow focus. In other words, the application of contrastive focus may influence the lexical tones of these new nouns, which increased the processing difficulty for English listeners.

Further, English and Mandarin listeners utilize prosodic information differently due to the influence of the L1 prosodic system. English listeners tend to rely more on local prosodic cues like duration and intensity (Scharenborg et al., 2020). This can cause more challenges in online L2 prosodic comprehension for English listeners than for Mandarin listeners, who rely more on distributed prosodic cues like F0 (Scharenborg et al., 2020). Further, Mandarin achieves contrastive focus by extending pitch range and prolonging duration, potentially distorting pronunciation of lexical tones (Shih, 2004). When the nouns were repeated, non-native listeners could handle this distortion, since the target word was already highly activated by the earlier mention; but when the nouns were new, the distorted tones may have increased processing difficulty. Therefore, due to the interaction between listeners' L1 backgrounds and the concurrent language context being processed,

the findings especially indicate that the utilization and integration involving the prosodic information was showed to be different by Mandarin and English listeners.

7.3.3 Conclusion

Therefore, the conclusion of the current eye-tracking experiment testing listeners' perceptual flexibility demonstrated during real-time speech processing and comprehension is twofold: (1) *influence of old and new information sentences*: native and non-native listeners exhibit divergent perceptual patterns when processing sentences containing old and new information; (2) *language-specific influence*: native and non-native listeners demonstrate divergent perceptual patterns under the L1- and L2-specific linguistic nature.

Overall, non-native listeners generally exhibit a similar level of perceptual flexibility as native listeners, often adopting perceptual strategies akin to native listeners when processing old information sentences, as the target nouns' old information in sentences did not pose much difficulty for listeners. However, their treatment of prosodic cues, influenced by their native language backgrounds, results in reduced perceptual flexibility for non-native listeners in their reduced ability to utilize and integrate cues, particularly evident when processing sentences containing new information. In summary, differences in perceptual flexibility between native and non-native listeners can be attributed to two main factors. Firstly, the processing of old and new information sentences emphasizes the greater difficulty posed by new information targets, especially for non-native listeners, leading to their tendency to abandon verb semantics and struggle to integrate multiple cues compared to native listeners. Secondly, the influence of the native language's linguistic system, particularly the language-specific prosodic system, such as Mandarin and English, leads to differing utilization of prosodic information due to the distinct ways of conveying prosody in each language.

Chapter 8

General discussion

8.1 Summary of findings: native and non-native listeners' perceptual flexibility and language-specific influence

This thesis employed monitoring and eye-tracking paradigms to explore the perceptual flexibility of human listeners. This exploration aimed to illustrate the reasons behind listeners' effortless performance in native speech perception and their difficulties with non-native speech perception. The results of these experiments highlight the differences in perceptual flexibility between native and non-native listeners, across both "low-level" speech unit perception and "high-level" sentence processing.

Overall, the findings are threefold.

First, during speech unit perception, a comparable degree of perceptual flexibility was observed between native and non-native listeners. In the Mandarin context, Mandarin listeners demonstrated a consistent masking effect and were more stable than English listeners in identifying both phonemes and syllables. Furthermore, under both English and Mandarin contexts, syllable processing was generally found to be more easily disrupted, particularly by the artificial accent, as indicated by the fluctuation in the size of the masking effect, which resulted from more changes in RTs for syllables than for phonemes. This disruption in syllable processing suggests that listeners spent more time processing syllables than phonemes, and therefore, syllables generally received more attention under the

bottom-up and top-down information manipulation. These findings further complement the unit retrieval and representation aspects of existing exemplar theories.

Second, during sentence processing and comprehension, native and non-native listeners began to show marked differences in their ability to utilize and integrate multiple cues, especially evident when processing sentences containing new information as the target object for detection.

Third, the findings reveal that perceptual flexibility is influenced by the linguistic characteristics and attributes specific to a language, as especially demonstrated in eye-tracking Experiments 3 and 4. Non-native listeners' perceptual strategies in L2 can be transferred from their L1, and the linguistic nature of L2 can also affect non-native listeners' processing. This divergence underscores the language-specific nature of perceptual flexibility and its impact on speech perception and comprehension.

In monitoring Experiments 1 and 2, perceptual flexibility was tested through whether listeners can flexibly shift attention between phonemes and syllables under the manipulation of information flow, as manifested by the fluctuation in the size of the masking effect demonstrated by both native and non-native listeners. There are three aspects to note. First, as outlined previously, especially in the English context, native and non-native listeners exhibited comparable perceptual flexibility in shifting their attention between syllables and phonemes. With more significant fluctuations in masking effects, native listeners showed slightly higher flexibility than non-native listeners. In the Mandarin context, non-native listeners demonstrated higher flexibility than native listeners by giving more attention to syllables. According to using exemplar models to highlight potential differences between native and non-native speech perception, these differences could be attributed to the feature vectors of native and non-native listeners, where detailed features or inaccurate features presented in the vector could lead to contradictions when matching the input and memory traces. Whether non-native listeners are more or less "sensitive" than native listeners may depend on the level of their L2 proficiency.

Second, listeners exhibited higher sensitivity to syllables than to phonemes under the

information flow manipulation. Syllables were found in Experiments 1 and 2 to be the primary focus for both native and non-native listeners.

Third, although the adaptive resonance in ART is the dynamic product of the interaction between bottom-up and top-down information, the main effect of bottom-up information found in Experiments 1 and 2 is greater than that of top-down information. This aligns with some previous studies outlined in Chapter 2 (Delphine Dahan, James S. Magnuson, 2006; Norris et al., 2000)), which argue that the top-down information flow is never necessary and is not as important as the bottom-up input in speech perception.

In the eye-tracking Experiment 3 and 4, the perceptual flexibility was tested through whether listeners can flexibly utilize and integrating multiple cues during the real-time sentence processing. The perceptual flexibility of native and non-native listeners showed drastic difference during the high-level sentence processing. Language-specific perceptual strategies also emerged when listeners processing sentences with target object as the new information to detect.

When listeners processed sentences with target object as the repeated information to detect, there was no language-specific influence, under both English and Mandarin context, native and non-native listeners adopted similar perceptual strategies. However, especially when processing the new information sentences, the native listeners did show higher perceptual flexibility by being sensitive to both prosodic and semantic cues and always showed more integration of these cues than non-native listeners. By contrast, when presented with new information, non-native listeners abandoned processing particular cues and showed less ability to integrate these cues.

Combining both low-level speech perception and high-level sentence processing, the perceptual flexibility demonstrated by native and non-native listeners differs. Native and non-native listeners showed comparable perceptual flexibility during low-level speech unit perception. However, differences in perceptual flexibility emerge during high-level sentence processing. Native listeners are generally more flexible in utilizing and integrating multiple cues compared to non-native listener.

8.2 Limitations

One methodological limitation of this thesis was the participant recruitment. First, English native listeners learning Mandarin as an L2 were not recruited from Chinese-speaking areas. Due to the COVID-19 travel restrictions from 2020 to 2023 in China, all participants had to be recruited in the UK. Ideally, fieldwork would have been conducted in Chinese-speaking environment at universities with eye-tracking equipment to recruit Mandarin L2 participants. This approach would have ensured that these Mandarin L2 participants had comparable proficiency levels to the English L2 participants recruited in the UK. It would be ideal to strictly control the proficiency of Mandarin L2 participants and ensure they use Mandarin every day in a Chinese-speaking environment.

Second, since all the participants were recruited in the UK, all Mandarin native listeners were Mandarin-English bilingual speakers. However, not all English native listeners were English-Mandarin bilinguals; some were English monolinguals or bilinguals of languages other than Mandarin. It would be ideal to select either all English-Mandarin bilinguals or all English/Mandarin monolinguals as native listeners in the experiments.

Further, the age of the listeners was not recorded, as most of the participants were university students recruited at the University of Glasgow, and their ages were controlled to be over 18. However, a few participants were over 40 years old, with their exact ages unknown.

Moreover, in the eye-tracking experiments, participants were asked to rate the sensibility of sentences by pressing a keyboard from "1" to "4," with "1" indicating "very sensible" and "4" indicating "very insensible." However, the data from these ratings was not successfully recorded, leaving the participants' judgments on the sentences' sensibility unknown.

8.3 Conclusion and future directions

The two overarching questions in the field of speech perception and comprehension have been asked (Pisoni, 2018): What makes native speech perception effortless and non-native speech perception difficult? This thesis has explored the concept of perceptual flexibility to address these inquiries, examining how such flexibility manifests itself in the domain of speech unit perception and in the domain of utilization and integration of cues.

As discussed in Section 5.1, various theories are presented to explain listeners' potentially differing performance during native and non-native speech perception and sentence processing, including factors such as L2 proficiency, L1 influences, cue preferences, cognitive load, and language-specific systems. These theories identify these factors influencing real-time native and non-native speech processing. They describe what listeners are equipped with and how external conditions shape processing. In contrast to these previous theories in the field of speech perception, the concept of perceptual flexibility introduced in this thesis represents a cognitive mechanism that explains listeners' potential differences in native and non-native speech processing at both the sub-lexical and sentence processing levels. This cognitive mechanism provides an internal, process-oriented explanation for listener behavior, with a focus on flexible attentional control. The factors mentioned in previous studies define the boundaries and conditions under which perceptual flexibility operates, shaping the extent to which listeners can exhibit flexibility. For example, higher proficiency in L2 may enable greater perceptual flexibility by providing listeners with a richer linguistic framework to guide flexible attention. In contrast, higher cognitive load may limit perceptual flexibility by constraining listeners' attentional resources. L1 influence may bias the starting point of attentional allocation, such as phoneme-biased or syllable-biased processing. Stimulus language that listeners are processing may influence how flexibility is exercised, given the linguistic-specific properties of each language.

Recall the results as previously discussed. During low-level speech unit perception, although native listeners under the Mandarin context did not show a preference for either phoneme or syllable under the information flow manipulation, native and non-native lis-

teners generally exhibited comparable perceptual flexibility in terms of shifting attention across perceptual units of different sizes. In the domain of high-level sentence processing, divergent perceptual flexibility emerged between native and non-native listeners when processing sentences containing target nouns as new information. Non-native listeners demonstrated reduced perceptual flexibility compared to native listeners by abandoning the utilization of certain cues and showing less capability in integrating multiple cues during the processing of sentences with new information. This finding suggests that human listeners may exhibit greater divergence in native and non-native speech perception at higher levels of speech processing and comprehension, rather than at the initial stages, such as low-level speech unit perception.

The findings of monitoring and eye-tracking experiments in this thesis raise unresolved questions and inspire further investigation into the domain of how different types of distortion affect syllable processing in different ways. Not all phoneme-level manipulations would necessarily affect processing of syllables as much, or in the same way. For example, in English, lengthening a lateral fricative, such as /ɬ/ or altering its quality may not significantly impact the entire syllable. However, retroflexion tends to "color" the adjacent vowel acoustically to a some extent. If not only the consonant sounds unexpected but the vowel also sounds "weird," these effects on the vowel could influence listeners' processing of the syllable. Similar situations could also occur with syllable-initial ejectives (a brief glottal stop as a "disruption" following the initial /k/-like sound) and the palatal nasal (a /j/-like sound following the /n/-like sound). Listeners may be robust to these disruptions in phoneme processing, but they could be significantly affected at the syllable level.

Moreover, this distortion effect on syllable processing could be language-specific. As demonstrated by Experiment 2, in the Mandarin context, neither native nor non-native listeners showed greater disruption in syllable processing compared to phoneme processing. With a consistent magnitude of masking effects, Mandarin listeners were more stable than English listeners in identifying both phonemes and syllables. This suggests that the sounds mentioned above, which could be disruptive at the syllable level in English, may

not be as disruptive at the Mandarin syllable level, especially for Mandarin native listeners when processing Mandarin.

Secondly, does contrastive pitch accent actually facilitate listeners' processing of new information structure, despite accurately signaling it? As demonstrated in Experiments 3 and 4, English listeners' detection of the target noun did not substantially benefit from the contrastive pitch accent applied to the new target noun in both English and Mandarin contexts, even though the accent accurately indicated its new information structure. According to previous findings, prosodic prominence on the target deepens semantic processing, eliciting a stronger N400 response for listeners (Brunellière et al., 2019; Ventura et al., 2020; L. Wang et al., 2011), and reduces the N400 response for the post-focus position (Ventura et al., 2020). As stated in Section 7.1 of Chapter 7, the effect of incongruence at the pre-focus position may also be attenuated by prosodic prominence at the focus position, as supported by results from the current eye-tracking experiment indicating that the negative effect of the inappropriate verb was lessened by the presence of the contrastive pitch accent on the target for English listeners. Therefore, further evidence is required to investigate whether prosodic prominence on the target noun distracts listeners from other linguistic cues at pre-focus and post-focus positions, thus exerting a negative effect on detection for listeners. Further, this could also be language-specific due to divergent prosodic systems, as Mandarin listeners significantly benefited from prosodic prominence on the target noun as new information for detection in both English and Mandarin contexts. Consequently, further investigation into semantic processing at pre-focus and post-focus positions in Mandarin is warranted. This inquiry aims to ascertain whether prosodic prominence at the focus position similarly diminishes, enhances, or remains indifferent to semantic processing of information at pre-focus and post-focus positions in Mandarin, which further sheds light on listeners' processing differences across languages and the impact of divergent prosodic systems.

In summary, despite lingering questions, this thesis has revealed that non-native listeners, although less perceptually flexible than native listeners in processing sentences and handling new information, can still flexibly shift their attention during speech unit percep-

tion and adjust their perceptual strategies during sentence processing. While non-native speech perception may present challenges, non-native listeners have access to the same perceptual strategies as native listeners, even if they are not always able to deploy them effectively in all contexts.

Bibliography

Akker, E., & Cutler, A. (2003). Prosodic cues to semantic structure in native and nonnative listening. *Bilingualism: Language and Cognition*, *6*(2), 81–96.

Altmann, G. T. M., & Kamide, Y. (2009). Discourse-mediation of the mapping between language and the visual world: Eye movements and mental representation. *Cognition*, *111*(1), 55–71.

Altmann, G. T. M., & Mirković, J. (2009). Incrementality and Prediction in Human Sentence Processing. *Cognitive Science*, *33*(4), 583–609.

Altmann, G. T., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–264.

Altmann, G. T., & Kamide, Y. (2007). The real-time mediation of visual attention by language and world knowledge: Linking anticipatory (and other) eye movements to linguistic processing. *Journal of Memory and Language*, *57*(4), 502–518.

Arai, M., & Keller, F. (2013). The use of verb-specific information for prediction in sentence processing. *Language and Cognitive Processes*, *28*(4), 525–560.

Balota, D. A., & Paul, S. T. (1996). Summation of activation: Evidence from multiple primes that converge and diverge within semantic memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(4), 827–845.

Baumann, S., & Schumacher, P. B. (2012). (De-)Accentuation and the processing of information status: Evidence from event-related brain potentials. *Language and Speech*, *55*(3), 361–381.

Bent, T., Baese-Berk, M., Borrie, S. A., & McKee, M. (2016). Individual differences in the perception of regional, nonnative, and disordered speech varieties. *The Journal of the Acoustical Society of America*, *140*(5), 3775–3786.

Bever, T. G., & Poeppel, D. (2010). Analysis by Synthesis: A (Re-)Emerging Program of Research for Language and Vision. *Biolinguistics*, *4*(2-3), 174–200.

Birch, S., & Clifton, C. (1995). Focus, Accent, and Argument Structure: Effects on Language Comprehension. *Language and Speech*, *38*(4), 365–391.

Blake, A., & Palmisano, S. (2021). Divergent Thinking Influences the Perception of Ambiguous Visual Illusions. *Perception*, *50*(5), 418–437.

Bock, J. K., & Mazzella, J. R. (1983). Intonational marking of given and new information: Some consequences for comprehension. *Memory & Cognition*, *11*(1), 64–76.

Bohn, O.-S. (2018). Cross-language and second language speech perception. In E. M. Fernández & H. S. Cairns (Eds.), *The Handbook of Psycholinguistics* (1st ed., pp. 214–239). John Wiley & Sons Inc.

Bowers, J. S., & Davis, C. J. (2004). Is speech perception modular or interactive? *Trends in Cognitive Sciences*, *8*(1), 3–5.

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, *106(2)*, 707–729.

Broersma, M., & Cutler, A. (2008). Phantom word activation in L2. *System*, *36*(1), 22–34.

Brown, M., & Kuperberg, G. R. (2015). A Hierarchical Generative Framework of Language Processing: Linking Language Perception, Interpretation, and Production Abnormalities in Schizophrenia. *Frontiers in Human Neuroscience*, *9*.

Brunellière, A., Auran, C., & Delrue, L. (2019). Does the prosodic emphasis of sentential context cause deeper lexical-semantic processing? *Language, Cognition and Neuroscience*, *34*(1), 29–42.

Carpenter, G. A., & Grossberg, S. (1987). ART 2: Self-organization of stable category recognition codes for analog input patterns [Publisher: Optica Publishing Group]. *Appl. Opt.*, *26*(23), 4919–4930.

Chambers, C. G., & Cooke, H. (2009). Lexical competition during second-language listening: Sentence context, but not proficiency, constrains interference from the native lexicon. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(4), 1029–1040.

Chambers, C. G., Tanenhaus, M. K., Eberhard, K. M., Filip, H., & Carlson, G. N. (2002). Circumscribing Referential Domains during Real-Time Language Comprehension. *Journal of Memory and Language*, *47*(1), 30–49.

Chen, Y., & Braun, B. (2006). Prosodic Realization of Information Structure Categories in Standard Chinese. *Proceedings of the 3rd International Conference on Speech Prosody*, 54: 5–8.

Choi, J., Cutler, A., & Broersma, M. (2017). Early development of abstract language knowledge: Evidence from perception–production transfer of birth-language memory. *Royal Society Open Science*, *4*(1), 160660.

Chrabaszcz, A., Winn, M., Lin, C. Y., & Idsardi, W. J. (2014). Acoustic Cues to Perception of Word Stress by English, Mandarin, and Russian Speakers. *Journal of Speech, Language, and Hearing Research*, *57*(4), 1468–1479.

Chun, E., & Kaan, E. (2019). L2 Prediction during complex sentence processing. *Journal of Cultural Cognitive Science*, *3*(2), 203–216.

Ciocca, V., & Bregman, A. S. (1989). The effects of auditory streaming on duplex perception. *Perception & Psychophysics*, *46*(1), 39–48.

Clahsen, H., & Felser, C. (2006a). Continuity and shallow structures in language processing. *Applied Psycholinguistics*, *27*(1), 107–126.

Clahsen, H., & Felser, C. (2006b). Grammatical processing in language learners. *Applied Psycholinguistics*, *27*(1), 3–42.

Clark, H. H., & Haviland, S. E. (1977). Comprehension and the Given-New Contract. In R. O. Freedle (Ed.), *Discourse Production and Comprehension*.

Cole, J. (2015). Prosody in context: A review. *Language, Cognition and Neuroscience*, *30*(1-2), 1–31.

Corps, R. E., Liao, M., & Pickering, M. J. (2023). Evidence for two stages of prediction in non-native speakers: A visual-world eye-tracking study. *Bilingualism: Language and Cognition*, *26*(1), 231–243.

Corps, R. E., & Rabagliati, H. (2020). How top-down processing enhances comprehension of noise-vocoded speech: Predictions about meaning are more important than predictions about form. *Journal of Memory and Language*, *113*, 104114.

Coughlin, C. E., & Tremblay, A. (2012). Non-Native Listeners' Delayed Use of Prosodic Cues in Speech Segmentation. *Proceedings of the Australasian International Conference on Speech Science and Technology*.

Cutler, A., Eisner, F., McQueen, J. M., & Norris, D. (2010). How abstract phonemic categories are necessary for coping with speaker-related variation. In C. Fougeron, B. Kühnert, & N. Vallée (Eds.), *Laboratory Phonology* (pp. 99–111). Berlin : De Gruyter Mouton.

Cutler, A. (1976). Phoneme-monitoring reaction time as a function of preceding intonation contour. *Perception & Psychophysics*, *20*(1), 55–60.

Cutler, A. (2008). The 34th Sir Frederick Bartlett Lecture: The abstract representations in speech processing [_eprint: https://doi.org/10.1080/13803390802218542]. *Quarterly Journal of Experimental Psychology*, *61*(11), 1601–1619.

Cutler, A. (2012). *Native Listening: Language Experience and the Recognition of Spoken Words*. The MIT Press.

Cutler, A. (2017). Converging Evidence for Abstract Phonological Knowledge in Speech Processing. *Annual Meeting of the Cognitive Science Society*.

Cutler, A., & Clifton, J., C. (1984). The use of prosodic information in word recognition. In H. Bouma & D. G. Bouwhuis (Eds.), *Attention and performance X: Control of language processes* (pp. 183–196).

Cutler, A., & Foss, D. J. (1977). On the Role of Sentence Stress in Sentence Processing. *Language and Speech*, *20*(1), 1–10.

Cutler, A., & Norris, D. (1988). The Role of Strong Syllables in Segmentation for Lexical Access. *Journal of Experimental Psychology: Human Perception and Performance*, *14(1)*, 113–121.

Cutler, A., Weber, A., & Otake, T. (2006). Asymmetric mapping from phonetic to lexical representations in second-language listening. *Journal of Phonetics*, *34*(2), 269–284.

Cutting, J. E. (1975). Aspects of phonological fusion. *Journal of Experimental Psychology: Human Perception and Performance*, *104*, 105–120.

Dahan, D., Tanenhaus, M. K., & Chambers, C. G. (2002). Accent and reference resolution in spoken-language comprehension. *Journal of Memory and Language*, *47*(2), 292–314.

Davis, M. H., & Johnsrude, I. S. (2007). Hearing speech sounds: Top-down influences on the interface between audition and speech perception. *Hearing Research*, *229*(1-2), 132–147.

Decoene, S. (1993). Testing the speech unit hypothesis with the primed matching task: Phoneme categories are perceptually basic. *Perception & Psychophysics*, *53*(6), 601–616.

Delphine Dahan, James S. Magnuson. (2006). Chapter 8 Spoken Word Recognition. In *Handbook of psycholinguistics* (2nd ed, pp. 249–283). Elsevier.

Diehl, R. L., & Kluender, K. R. (1989). On the objects of speech perception. *Ecological Psychology*, *1*(2), 121–144.

Dienes, Z., Altmann, G. T. M., & Gao, S.-J. (1999). Mapping across domains without feedback: A neural network model of transfer of implicit knowledge. *Cognitive Science*, *23*(1), 53–82.

Dijkgraaf, A., Hartsuiker, R. J., & Duyck, W. (2017). Predicting upcoming information in native-language and non-native-language auditory word recognition. *Bilingualism: Language and Cognition*, *20*(5), 917–930.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211.

Elman, J. L., & McClelland, J. L. (1986). Exploiting Lawful Variability In the Speech Wave. In J. S. Perkell & D. H. Klatt (Eds.), *Invariance and Variability in Speech Processes* (1st Edition, pp. 360–385). Psychology Press.

Ezzatian, P., Avivi, M., & Schneider, B. A. (2010). Do nonnative listeners benefit as much as native listeners from spatial cues that release speech from masking? *Speech Communication*, *52*(11), 919–929.

FitzPatrick, I., & Indefrey, P. (2010). Lexical Competition in Nonnative Speech Comprehension. *Journal of Cognitive Neuroscience*, *22*(6), 1165–1178.

Foltz, A. (2021). Using prosody to predict upcoming referents in the L1 and L2. *Studies in Second Language Acquisition*, *43*(4), 753–780.

Foss, D. J., & Swinney, D. A. (1973). On the psychological reality of the phoneme: Perception, identification, and consciousness. *Journal of Verbal Learning & Verbal Behavior*, *12*(3), 246–257.

Foss, D. (1971). On the time-course of sentence comprehension. *The symposium on Les Problems Actuels de Psycholinguistique, Centre National de Recherche Scientifique, Paris*.

Foucart, A., Ruiz-Tada, E., & Costa, A. (2016). Anticipation processes in L2 speech comprehension: Evidence from ERPs and lexical recognition task. *Bilingualism: Language and Cognition*, *19*(1), 213–219.

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct–realist perspective. *Journal of Phonetics*, *14*(1), 3–28.

Francis, A. L., Ciocca, V., Ma, L., & Fenn, K. (2008). Perceptual learning of Cantonese lexical tones by tone and non-tone language speakers. [Place: Netherlands Publisher: Elsevier Science]. *Journal of Phonetics*, *36*(2), 268–294.

Frank, S. L., & Willems, R. M. (2017). Word predictability and semantic similarity show distinct patterns of brain activity during language comprehension. *Language, Cognition and Neuroscience*, *32*(9), 1192–1203.

Frauenfelder, U. H., & Segui, J. (1989). Phoneme monitoring and lexical processing: Evidence for associative context effects. *Memory & Cognition*, *17*(2), 134–140.

Freeman, M. R., Blumenfeld, H. K., Carlson, M. T., & Marian, V. (2022). First-language influence on second language speech perception depends on task demands. *Language and Speech*, *65*(1), 28–51.

Frixione, M., & Lieto, A. (2012). Representing Concepts in Formal Ontologies: Compositionality vs. Typicality Effects". *Logic and Logical Philosophy*, *21*(4), 391–414.

Ganga, R., Ge, H., Struiksma, M. E., Yip, V., & Chen, A. (2024). Prosodic processing in sentences with 'only' in L1 and L2 English. *Studies in Second Language Acquisition*, *46*(2), 478–503.

Gaskell, M. G., & Marslen-Wilson, W. D. (1997). Integrating Form and Meaning: A Distributed Model of Speech Perception. *Language and Cognitive Processes*, *12*(5-6), 613–656.

Goldinger, S. D. (1998). Echoes of Echoes? An Episodic Theory of Lexical Access. *Psychological Review*, *Vol 105. No. 2*, 251–279.

Goldinger, S. D. (2007). A complementary-systems approach to abstract and episodic speech perception. *Proceedings of the 16th International Congress of Phonetic Sciences*, 49–54.

Goldinger, S. D., & Azuma, T. (2003). Puzzle-solving science: The quixotic quest for units in speech perception. *Journal of Phonetics*, *31*(3-4), 305–320.

Goldinger, S. D., & Azuma, T. (2004). Resonance within and between linguistic beings. *Behavioral and Brain Sciences*, *27*(2), 199–200.

Goldstein, L., & Fowler, C. A. (2003). Articulatory Phonology: A phonology for public language use. In N. O. Schiller & A. S. Meyer (Eds.), *Phonetics and Phonology in Language Comprehension and Production* (pp. 159–208). DE GRUYTER MOUTON.

Goldstone, R. L. (1998). PERCEPTUAL LEARNING. *Annual Review of Psychology*, *49*(1), 585–612.

Grossberg, S. (1999). The Link between Brain Learning, Attention, and Consciousness. *Consciousness and Cognition*, *8*, 44.

Grossberg, S. (1976a). Adaptive pattern classification and universal recoding, i: Parallel development and coding of neural feature detectors. *Biological Cybernetics*, *23*, 121–134.

Grossberg, S. (1976b). Adaptive pattern classification and universal recoding: II. Feedback, expectation, olfaction, illusions. *Biological Cybernetics*, *23*(4), 187–202.

Grossberg, S. (1980). How does a brain build a cognitive code? *Psychological Review*, *87*(1), 1–51.

Grossberg, S. (2003). The Brain's Cognitive Dynamics: The Link between Learning, Attention, Recognition, and Consciousness. In V. Palade, R. J. Howlett, & L. Jain (Eds.), *Knowledge-Based Intelligent Information and Engineering Systems* (pp. 5–12). Springer Berlin Heidelberg.

Grossberg, S. (2013). Adaptive Resonance Theory: How a brain learns to consciously attend, learn, and recognize a changing world. *Neural Networks*, *37*, 1–47.

Grossberg, S., Boardman, I., & Cohen, M. (1997). Neural dynamics of variable-rate speech categorization. *Journal of Experimental Psychology: Human Perception and Performance*, *23*, 481–503.

Grossberg, S., & Kazerounian, S. (2016). Phoneme restoration and empirical coverage of Interactive Activation and Adaptive Resonance models of human speech processing. *The Journal of the Acoustical Society of America*, *140*(2), 1130–1153.

Gundel, J. K. (2012). Pragmatics and information structure. In K. Allan & K. M. Jaszczolt (Eds.), *The Cambridge Handbook of Pragmatics* (1st ed., pp. 585–598). Cambridge University Press.

Gundel, J. K., & Fretheim, T. (2006). Topic and Focus. In L. R. Horn & G. Ward (Eds.), *The Handbook of Pragmatics* (1st ed., pp. 175–196). Wiley.

Gwilliams, L., & Davis, M. (2020). Extracting language content from speech sounds: An information theoretic approach, 17.

Hahne, A., & Friederici, A. D. (2001). Processing a second language: Late learners' comprehension mechanisms as revealed by event-related brain potentials. *Bilingualism: Language and Cognition*, *4*(2), 123–141.

Halliday, M. A. K. (1967). Notes on transitivity and theme in English. *Journal of Linguistics*, *3*(1), 37–81.

Hannagan, T., Magnuson, J. S., & Grainger, J. (2013). Spoken word recognition without a TRACE. *Frontiers in Psychology*, *4*.

Hanulíková, A., Mitterer, H., & Mcqueen, J. M. (2011). Effects of first and second language on segmentation of non-native speech. *Bilingualism: Language and Cognition*, *14*(4), 506–521.

Heald, S. L. M., & Nusbaum, H. C. (2014). Speech perception as an active cognitive process. *Frontiers in Systems Neuroscience*, *8*.

Healy, A. F., & Cutting, J. E. (1976). Units of speech perception: Phoneme and syllable. *Journal of Verbal Learning and Verbal Behavior*, *15*(1), 73–83.

Hintzman, D. L. (1984). MINERVA 2: A simulation model of human memory. *Behavior Research Methods, Instruments, &amp Computers*, *16*(2), 96–101.

Hintzman, D. L. (1986). "Schema Abstraction" in a Multiple-Trace Memory Model. *Psychological Review*, *93*, 411–428.

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review*, *95*(4), 528–551.

Hockett, C. F. (1956). A Manual of Phonology. *International journal of American linguistics*, *Vol. 21, No. 4, Part 1*, 675–705.

Hodgson, J. M. (1991). Informational constraints on pre-lexical priming. *Language and Cognitive Processes*, *6*(3), 169–205.

Hopp, H. (2010). Ultimate attainment in L2 inflection: Performance similarities between non-native and native speakers. *Lingua*, *120*(4), 901–931.

Hopp, H. (2015). Semantics and morphosyntax in predictive L2 sentence processing. *International Review of Applied Linguistics in Language Teaching*, *53*(3), 277–306.

Huettig, F., Rommers, J., & Meyer, A. S. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta Psychologica*, *137*(2), 151–171.

Ingvalson, E. M., Holt, L. L., & McClelland, J. L. (2012). Can native Japanese listeners learn to differentiate /r–l/ on the basis of F3 onset frequency? – CORRIGENDUM. *Bilingualism: Language and Cognition*, *15*(2), 434–435.

Ip, M. H. K., & Cutler, A. (2020). Universals of listening: Equivalent prosodic entrainment in tone and non-tone languages. *Cognition*, *202*, 104311.

Ito, A., Pickering, M. J., & Corley, M. (2018). Investigating the time-course of phonological prediction in native and non-native speakers of English: A visual world eye-tracking study. *Journal of Memory and Language*, *98*, 1–11.

Ito, A., & Sakai, H. (2021). Everyday Language Exposure Shapes Prediction of Specific Words in Listening Comprehension: A Visual World Eye-Tracking Study. *Frontiers in Psychology*, *12*, 607474.

Ito, K., Bibyk, S. A., Wagner, L., & Speer, S. R. (2014). Interpretation of contrastive pitch accent in six- to eleven-year-old English-speaking children (and adults). *Journal of Child Language*, *41*(1), 84–110.

Ito, K., & Speer, S. R. (2008). Anticipatory effects of intonation: Eye movements during instructed visual search. *Journal of Memory and Language*, (58(2)), 541–573.

Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Tohkura, Y., Kettermann, A., & Siebert, C. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition*, *87*(1), B47–B57.

Jiang, N. (2007). Selective Integration of Linguistic Knowledge in Adult Second Language Learning. *Language Learning*, *57*(1), 1–33.

Johnson, E. K., Westrek, E., Nazzi, T., & Cutler, A. (2011). Infant ability to tell voices apart rests on language experience. *Developmental Science*, *14*(5), 1002–1011.

Johnson, K. (1997). Speech Perception without Speaker Normalization. In *Talker Variability in Speech Processing* (pp. 145–162). Academic Press Inc.

Johnson, K. (2005). Decisions and Mechanisms in Exemplar-based Phonology. *UC Berkeley Phonology Lab Annual Reports*, *1*.

Juffs, A., & Harrington, M. (1995). Parsing Effects in Second Language Sentence Processing: Subject and Object Asymmetries in wh-Extraction. *Studies in Second Language Acquisition*, *17*(4), 483–516.

Jusczyk, P. W., Friederici, A. D., Wessels, J. M. I., Svenkerud, V. Y., & Jusczyk, A. M. (1993). Infants Sensitivity to the Sound Patterns of Native Language Words. *Journal of Memory and Language*, *32*(3), 402–420.

Jusczyk, P. W., Luce, P. A., & Charles-Luce, J. (1994). Infants Sensitivity to Phonotactic Patterns in the Native Language. *Journal of Memory and Language*, *33*(5), 630–645.

Kaan, E. (2014). Predictive sentence processing in L2 and L1: What is different? *Linguistic Approaches to Bilingualism*, *4*, 257–282.

Kamide, Y. (2008). Anticipatory Processes in Sentence Processing. *Language and Linguistics Compass*, *2*(4), 647–670.

Kamide, Y., Altmann, G. T., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language*, *49*(1), 133–156.

Kamide, Y., Scheepers, C., & Altmann, G. T. M. (2003). Integration of Syntactic and Semantic Information in Predictive Processing: Cross-Linguistic Evidence from German and English. *Journal of Psycholinguistic Research*, *32*, 37–55.

Kilman, L., Zekveld, A., Hällgren, M., & Rönnberg, J. (2014). The influence of non-native language proficiency on speech perception performance. *Frontiers in Psychology*, *5*, Article 651.

Klatt, D. H. (1979). Speech perception: A model of acoustic-phonetic analysis and lexical access. *Journal of Phonetics*, *7*(3), 279–312.

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, *122*(2), 148–203.

Knoeferle, P., & Crocker, M. W. (2007). The influence of recent scene events on spoken comprehension: Evidence from eye movements. *Journal of Memory and Language*, *57*(4), 519–543.

Kraljic, T., & Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology*, *51*, 141–178.

Kurumada, C., Brown, M., Bibyk, S., Pontillo, D. F., & Tanenhaus, M. K. (2014). Is it or isn't it: Listeners make rapid use of prosody to infer speaker meanings. *Cognition*, *133*(2), 335–342.

Ladd, D. R., Mennen, I., & Schepman, A. (2000). Phonological conditioning of peak alignment in rising pitch accents in Dutch. *The Journal of the Acoustical Society of America*, *107*(5), 2685–2696.

Lagrou, E., Hartsuiker, R. J., & Duyck, W. (2013). The influence of sentence context and accented speech on lexical access in second-language auditory word recognition. *Bilingualism: Language and Cognition*, *16*(3), 508–517.

Landauer, T., & Streeter, L. (1973). Structural differences between common and rare words: Failure of equivalence assumptions for theories of word recognition. *Journal of Verbal Learning and Verbal Behavior*, *12*(2), 119–131.

Laufer, I., & Pratt, H. (2003). Evoked potentials to auditory movement sensation in duplex perception. *Clinical Neurophysiology*, *114*(7), 1316–1331.

Lee, S., Kang, J., & Nam, H. (2022). Identification of English vowels by non-native listeners: Effects of listeners' experience of the target dialect and talkers' language background. *Second Language Research*, *38*(3), 449–475.

Lee, Y.-C., Wang, T., & Liberman, M. (2016). Production and Perception of Tone 3 Focus in Mandarin Chinese. *Frontiers in Psychology*, *7*.

Lew-Williams, C., & Fernald, A. (2007). How First and Second Language Learners Use Predictive Cues in Online Sentence Interpretation in Spanish and English. *Proceedings of the 31st Annual Boston University Conference on Language Development*, *2*, 382–393.

Li, X., & Lu, Y. (2011). How accentuation influences semantic short-term memory representations during online speech processing: An event-related potential study. *Neuroscience*, *193*, 217–228.

Li, X., & Ren, G. (2012). How and when accentuation influences temporarily selective attention and subsequent semantic processing during online spoken language comprehension: An ERP study. *Neuropsychologia*, *50*(8), 1882–1894.

Liberman, A., & Whalen, D. (2000). On the relation of speech to language. *Trends in Cognitive Sciences*, *4*(5), 10.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*(6), 431–461.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, *21*(1), 1–36.

Liu, F., & Xu, Y. (2005). Parallel Encoding of Focus and Interrogative Meaning in Mandarin Intonation. *Phonetica*, *62*(2-4), 70–87.

Liu, Y., & Ning, J. (2018). The perception of Mandarin focus intonation by native English speakers. *6th International Symposium on Tonal Aspects of Languages (TAL 2018)*, 232–236.

Llanos, F., German, J. S., Gnanateja, G. N., & Chandrasekaran, B. (2021). The neural processing of pitch accents in continuous speech. *Neuropsychologia*, *158*, 107883.

Lorenz, D., & Tizón-Couto, D. (2019). Chunking or predicting – frequency information and reduction in the perception of multi-word sequences. *Cognitive Linguistics*, *30*(4), 751–784.

Luce, P. A., & Large, N. R. (2001). Phonotactics, density, and entropy in spoken word recognition. *Language and Cognitive Processes*, *16*(5-6), 565–581.

Magnuson, J. S., Mirman, D., Luthra, S., Strauss, T., & Harris, H. D. (2018). Interaction in Spoken Word Recognition Models: Feedback Helps. *Frontiers in Psychology*, *9*, 369.

Mann, V. A., & Liberman, A. M. (1983). Some differences between phonetic and auditory modes of perception. *Cognition*, *14*(00100277), 211–235.

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition*, *25*(1-2), 71–102.

Massaro, D. W. (1974). Perceptual units in speech recognition. *Journal of Experimental Psychology*, *102*(2), 199–208.

Massaro, D. W. (1989). Testing between the TRACE model and the fuzzy logical model of speech perception. *Cognitive Psychology*, *21*(3), 398–421.

Massaro, D. W., & Chen, T. H. (2008). The motor theory of speech perception revisited. *Psychonomic Bulletin & Review*, *15*(2), 453–457.

Massaro, D. (1975). Preperceptual Images, Processing Time, and Perceptual Units in Speech Perception. In *Understanding Language* (pp. 125–150). Elsevier.

Mattingly, I. (1990). Speech and other auditory modules. In *Signal and Sense: Local and Global Order in Perceptual Maps* (pp. 501–519). John Wiley & Sons.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*(1), 1–86.

McClelland, J. L., Mirman, D., Bolger, D. J., & Khaitan, P. (2014). Interactive activation and mutual constraint satisfaction in perception and cognition. *Cognitive Science*, *38*, 1139–1189.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological Review*, *88*, 375–407.

McDonald, J. L. (2006). Beyond the critical period: Processing-based explanations for poor grammaticality judgment performance by late second language learners. *Journal of Memory and Language*, *55*(3), 381–401.

McManus, I. C., Freegard, M., Moore, J., & Rawles, R. (2010). Science in the Making: Right Hand, Left Hand. II: The duck–rabbit figure. *Laterality: Asymmetries of Body, Brain and Cognition*, *15*(1-2), 166–185.

McNeill, D., & Lindig, K. (1973). The perceptual reality of phonemes, syllables, words, and sentences. *Journal of Verbal Learning & Verbal Behavior*, *12*, 419–430.

McQueen, J. M., & Cutler, A. (2010). Cognitive Processes in Speech Perception. In *The Handbook of Phonetic Sciences* (pp. 489–520). Jon Wiley & sons, Ltd.

McQueen, J. M., Cutler, A., & Norris, D. (2006). Phonological Abstraction in the Mental Lexicon. *Cognitive Science*, *30*(6), 1113–1126.

Mehler, J., Dommergues, J. Y., Frauenfelder, U., & Segui, J. (1981). The syllable's role in speech segmentation. *Journal of Verbal Learning and Verbal Behavior*, *20*(3), 298–305.

Meltzer, J. A., & Braun, A. R. (2013). P600-like positivity and Left Anterior Negativity responses are elicited by semantic reversibility in nonanomalous sentences. *Journal of neurolinguistics*, *26*(1), 10.1016/j.jneuroling.2012.06.001.

Meyer, L. (2018). The neural oscillations of speech processing and language comprehension: State of the art and emerging mechanisms. *European Journal of Neuroscience*, *48*(7), 2609–2621.

Miller, G. (1962). Decision units in the perception of speech. *IEEE Transactions on Information Theory*, *8*(2), 81–83.

Mirman, D. (2008). Mechanisms of Semantic Ambiguity Resolution: Insights from Speech Perception. *Research on Language and Computation*, *6*(3-4), 293–309.

Mirman, D., McClelland, J. L., & Holt, L. L. (2006). An interactive Hebbian account of lexically guided tuning of speech perception. *Psychonomic Bulletin & Review*, *13*(6), 958–965.

Mitsugi, S., & Macwhinney, B. (2016). The use of case marking for predictive processing in second language Japanese. *Bilingualism: Language and Cognition*, *19*(1), 19–35.

Morett, L. M., Fraundorf, S. H., & McPartland, J. C. (2021). Eye see what you're saying: Contrastive use of beat gesture and pitch accent affects online interpretation of spoken discourse. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *47*(9), 1494–1526.

Most, R. B., & Saltz, E. (1979). Information Structure in Sentences: New Information. *Language and Speech*, *22*(1), 89–95.

Nakano, Y., Felser, C., & Clahsen, H. (2002). Antecedent Priming at Trace Positions in Japanese Long-Distance Scrambling. *Journal of Psycholinguistic Research*, *31*, 531–571.

Neely, J. H. (1990). Semantic Priming Effects in Visual Word Recognition: A Selective Review of Current Findings and Theories. In D. Besner & G. W. Humphreys (Eds.), *Basic Processes in Reading: Visual Word Recognition* (pp. 264–336). Lawrence Erlbaum Associates.

Nooteboom, S. G., & Kruyt, J. G. (1984, December). Acceptability of Accenting and De-accenting 'NEW' and 'GIVEN' in Dutch. In A. Cohen & M. P. R. V. D. Broecke (Eds.), *Proceedings of the Tenth International Congress of Phonetic Sciences* (pp. 549–553). Boston:De Gruyter Mouton.

Nooteboom, S. G., & Kruyt, J. G. (1987). Accents, focus distribution, and the perceived distribution of given and new information: An experiment. *The Journal of the Acoustical Society of America*, *81*(S1), S67–S67.

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition*, *52*(3), 189–234.

Norris, D., & Cutler, A. (1988). The relative accessibility of phonemes and syllables. *Perception & Psychophysics*, *43*(6), 541–550.

Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences*, *23*(3), 299–325.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*(2), 204–238.

Nosofsky, R. M. (1992). Similarity Scaling and Cognitive Process Models, 29.

Nosofsky, R. M. (1986). Attention, Similarity, and the Identification-Categorization Relationship, 19.

Nosofsky, R. M. (1988). Similarity, Frequency, and Category Representations. *Journal of Experimental Psychology*, *14 No.1*, 54–65.

Nosofsky, R. M. (1991). Tests of an Exemplar Model for Relating Perceptual Classification and Recognition Memory. *Journal of Experimental Psychology*, *17. NO.1*, 3–27.

Nosofsky, R. M. (2011, January). The generalized context model: An exemplar model of classification. In E. M. Pothos & A. J. Wills (Eds.), *Formal Approaches in Categorization* (1st ed., pp. 18–39). Cambridge University Press.

Nosofsky, R. M., & Johansen, M. K. (2000). Exemplar-based accounts of "multiple-system" phenomena in perceptual categorization. *Psychonomic Bulletin and Review*, *7*(3), 375–402.

O'Brien, M. G., Jackson, C. N., & Gardner, C. E. (2014). Cross-linguistic differences in prosodic cues to syntactic disambiguation in German and English. *Applied Psycholinguistics*, *35*(1), 27–70.

Oden, G. C., & Massaro, D. W. (1978). Integration of featural information in speech perception. *Psychological Review*, *85*(3), 172–191.

Ouyang, I. C., & Kaiser, E. (2015). Prosody and information structure in a tone language: An investigation of Mandarin Chinese. *Language, Cognition and Neuroscience*, *30*(1-2), 57–72.

Patterson, C., Esaulova, Y., & Felser, C. (2017). The impact of focus on pronoun resolution in native and non-native sentence comprehension. *Second Language Research*, *33*(4), 403–429.

Pennington, M. C., & Ellis, N. C. (2000). Cantonese Speakers' Memory for English Sentences with Prosodic Cues. *The Modern Language Journal*, *84*(3), 372–389.

Perdomo, M., & Kaan, E. (2021). Prosodic cues in second-language speech processing: A visual world eye-tracking study. *Second Language Research*, *37*(2), 349–375.

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, *36*(4), 329–347.

Pierrehumbert, J. B., & Hirschberg, J. (1990). The Meaning of Intonational Contours in the Interpretation of Discourse. In P. R. Cohen, J. Morgan, & M. E. Pollack (Eds.), *Intentions in Communication* (pp. 271–311). Bradford Books, MIT Press.

Pintér, G., Mizuguchi, S., & Tateishi, K. (2014). Perception of prosodic prominence and boundaries by L1 and L2 speakers of English. *Interspeech 2014*, 544–547.

Pisoni, D. B. (1981). Some current theoretical issues in speech perception. *Cognition*, *10*(1-3), 249–259.

Pisoni, D. B. (2018). Speech perception: Research, theory, and clinical application. In *Handbook of psycholinguistics* (E.M. Fernández and H.S. Cairns). John Wiley & Sons Inc.

Pisoni, D. B. (1997). Some thoughts on "normalization" in speech perception. In *Talker Variability in Speech Processing* (K. Johnson & J.W. Mullennix, pp. 9–32). Academic Press.

Pitt, M. A., & Samuel, A. G. (1995). Lexical and Sublexical Feedback in Auditory Word Recognition. *Cognitive Psychology*, *29*(2), 149–188.

Qin, Z., Chien, Y.-F., & Tremblay, A. (2017). Processing of word-level stress by Mandarin-speaking second language learners of English. *Applied Psycholinguistics*, *38*(3), 541–570.

RCoreTeam. (2022). R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing. Vienna, Austria*. https://www.R-project.org/

Reed, M., & Michaud, C. (2015). Intonation in Research and Practice. In *The Handbook of English Pronunciation* (pp. 454–470). John Wiley & Sons, Ltd.

Roessig, S., Winter, B., & Mücke, D. (2022). Tracing the Phonetic Space of Prosodic Focus Marking. *Frontiers in Artificial Intelligence*, *5*, 842546.

Rosenberg, A., Hirschberg, J., & Manis, K. (2010). Perception of English Prominence by Native Mandarin Chinese Speakers. *Proceedings of the Fifth International Conference on Speech Prosody (SPro-2010)*, paper 982.

Saha, S. N., & Mandal, S. K. D. (2018). Phonetic realization of English lexical stress by native (L1) Bengali speakers compared to native (L1) English speakers. *Computer Speech & Language*, *47*, 1–15.

Samuel, A. G. (1982). Phonetic prototypes. *Perception & Psychophysics*, *31*(4), 307–314.

Samuel, A., & Kraljic, T. (2009). Perceptual learning for speech. *Attention, Perception, & Psychophysics*, *71*(6), 1207–1218.

Sanders, L. D., Neville, H. J., & Woldorff, M. G. (2002). Speech Segmentation by Native and Non-Native Speakers: The Use of Lexical, Syntactic, and Stress-Pattern Cues. *Journal of Speech, Language, and Hearing Research*, *45*(3), 519–530.

Savin, H. B., & Bever, T. G. (1970). The Nonperceptual Reality of the Phoneme I. *Journal of Verbal Learning and Verbal Behavior*, (9), 8.

Sawusch, J. R., & Gagnon, D. A. (1995). Auditory coding, cues, and coherence in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(3), 635–652.

Scharenborg, O., Kakouros, S., Post, B., & Meunier, F. (2020). Cross-linguistic Influences on Sentence Accent Detection in Background Noise. *Language and Speech*, *63*(1), 3–30.

Scharenborg, O., Kolkman, E., Kakouros, S., & Post, B. (2016). The Effect of Sentence Accent on Non-Native Speech Perception in Noise. *Interspeech 2016*, 863–867.

Schwartzhaupt, B., Alves, U. K., & Fontes, A. B. A. D. L. (2015). The role of L1 knowledge on L2 speech perception: Investigating how native speakers and Brazilian learners categorize different VOT patterns in English. *Revista de Estudos da Linguagem*, *23*(2), 311.

Sedivy, J. C., K. Tanenhaus, M., Chambers, C. G., & Carlson, G. N. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition*, *71*(2), 109–147.

Segalowitz, N. (2003). Automaticity and Second Languages. In C. J. Doughty & M. H. Long (Eds.), *The Handbook of Second Language Acquisition* (1st ed., pp. 382–408). Wiley.

Segui, J., Frauenfelder, U., & Mehler, J. (1981). Phoneme monitoring, syllable monitoring and lexical access. *British Journal of Psychology*, *72*(4), 471–477.

Selkirk, E. (1995). Sentence prosody: Intonation, stress and phrasing. In J. A. Goldsmith (Ed.), *The handbook of phonological theory* (pp. 550–569).

Shannon, C. E. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal*, *Vol. 27*, 379–423, 623–656.

Shepard, R. N. (1987). Toward a Universal Law of Generalization for Psychological Science. *Science*, *237*(4820), 1317–1323.

Shih, C. (2004). Tonal Effects on Intonation. *Proceedings of Interna- tional Symposium on Tonal Aspects of Languages: Emphasis on Tone Lan- guages*, 63–168.

Shuai, L., & Gong, T. (2014). Temporal relation between top-down and bottom-up processing in lexical tone perception. *Frontiers in Behavioral Neuroscience*, *8*.

Siegelman, N., Kearns, D. M., & Rueckl, J. G. (2020). Using information-theoretic measures to characterize the structure of the writing system: The case of orthographic-phonological regularities in English. *Behavior Research Methods*, *52*(3), 1292–1312.

Singh, L., & Chee, M. (2016). Rise and fall: Effects of tone and intonation on spoken word recognition in early childhood. *Journal of Phonetics*, *55*, 109–118.

Singh, L., Loh, D., & Xiao, N. G. (2017). Bilingual infants demonstrate perceptual flexibility in phoneme discrimination but perceptual constraint in face discrimination. *Frontiers in Psychology*, *8*.

Sjerps, M. J., & McQueen, J. M. (2010). The bounds on flexibility in speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *36*, 195–211.

Sohoglu, E., Peelle, J. E., Carlyon, R. P., & Davis, M. H. (2012). Predictive Top-Down Integration of Prior Knowledge during Speech Perception. *Journal of Neuroscience*, *32*(25), 8443–8453.

Sorace, A. (2011). Pinning down the concept of "interface" in bilinguals. *Linguistic Approaches to Bilingualism*, *1*, 1–33.

Sorace, A., & Filiaci, F. (2006). Anaphora resolution in near-native speakers of Italian. *Second Language Research*, *22*(3), 339–368.

Sóskuthy, M. (2017). Generalised additive mixed models for dynamic analysis in linguistics: A practical introduction [arXiv:1703.05339 [stat]].

Stanovich, K. E., & West, R. F. (1983). On priming by a sentence context. *Journal of Experimental Psychology: General*, *112*, 1–36.

Sun, C. C., Hendrix, P., Ma, J., & Baayen, R. H. (2018). Chinese lexical database (CLD): A large-scale lexical database for simplified Mandarin Chinese. *Behavior Research Methods*, *50*(6), 2606–2629.

Swinney, D. (1979). Lexical access during sentence comprehension: (Re)consideration of context effects. *Journal of Verbal Learning and Verbal Behavior*, *18*, 645–659.

Swinney, D. A., & Prather, P. (1980). Phonemic identification in a phoneme monitoring experiment: The variable role of uncertainty about vowel contexts. *Perception & Psychophysics*, *27*(2), 104–110.

Szewczyk, J. M., Mech, E. N., & Federmeier, K. D. (2022). The power of "good": Can adjectives rapidly decrease as well as increase the availability of the upcoming noun? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *48*(6), 856–875.

Tabri, D., Abou Chacra, K., & Pring, T. (2010). Speech perception in noise by monolingual, bilingual and trilingual listeners. *International Journal of Language & Communication Disorders*, *0*(0), 1–12.

Takahashi, C., Kao, S., Baek, H., Yeung, A. H., Hwang, J., & Broselow, E. I. (2018). Native and nonnative speaker processing and production of contrastive focus prosody. *Proceedings of the Linguistic Society of America*, *3*, 1–13.

Tang, P., Yuen, I., Demuth, K., & Rattanasone, N. X. (2023). The acquisition of contrastive focus during online sentence-comprehension by children learning Mandarin Chinese. *Developmental Psychology*, *59*(5), 845–861.

Terken, J., & Nooteboom, S. G. (1987). Opposite effects of accentuation and deaccentuation on verification latencies for given and new information. *Language and Cognitive Processes*, *2*(3-4), 145–163.

Theres Grüter, E. L., & Ling, W. (2020). How classifiers facilitate predictive processing in L1 and L2 Chinese: The role of semantic and grammatical cues. *Language, Cognition and Neuroscience*, *35*(2), 221–234.

Tremblay, A., Broersma, M., & Coughlin, C. E. (2018). The functional weight of a prosodic cue in the native language predicts the learning of speech segmentation in a second language. *Bilingualism: Language and Cognition*, *21*(3), anne.

Tremblay, A., Coughlin, C. E., Bahler, C., & Gaillard, S. (2012). Differential contribution of prosodic cues in the native and non-native segmentation of French speech. *Laboratory Phonology*, *3*(2).

Trofimovich, P., & Baker, W. (2007). Learning prosody and fluency characteristics of second language speech: The effect of experience on child learners' acquisition of five suprasegmentals. *Applied Psycholinguistics*, *28*(2), 251–276.

Tsurutani, C., Tsukada, K., & Ishihara, S. (2010). Comparison of Native and Non-native Perception of L2 Japanese Speech Varying in Prosodic Characteristics, 122–125.

Twilley, L. C., & Dixon, P. (2000). Meaning resolution processes for words: A parallel independent model. *Psychonomic Bulletin & Review*, *7*(1), 49–82.

Tyler, M. D., & Cutler, A. (2009). Cross-language differences in cue use for speech segmentation. *The Journal of the Acoustical Society of America*, *126*(1), 367–376.

Van Bergen, G., & Flecken, M. (2017). Putting things in new places: Linguistic experience modulates the predictive power of placement verb semantics. *Journal of Memory and Language*, *92*, 26–42.

Vanlancker–Sidtis, D. (2003). Auditory recognition of idioms by native and nonnative speakers of English: It takes one to know one. *Applied Psycholinguistics*, *24*(1), 45–57.

Vanpaemel, W., & Lee, M. D. (2012). Using priors to formalize theory: Optimal attention and the generalized context model. *Psychonomic Bulletin & Review*, *19*(6), 1047–1056.

van Rij, J., Wieling, M., Baayen, R. H., & van Rijn, H. (2022). Itsadug: Interpreting Time Series and Autocorrelated Data Using GAMMs.

Ventura, C., Grice, M., Savino, M., Kolev, D., Brilmayer, I., & Schumacher, P. B. (2020). Attention allocation in a language with post-focal prominences. *NeuroReport*, *31*(8), 624–628.

Vitevitch, M. S., & Luce, P. A. (1999). Probabilistic Phonotactics and Neighborhood Activation in Spoken Word Recognition. *Journal of Memory and Language*, *40*(3), 374–408.

Vitevitch, M. S., & Luce, P. A. (2004). A Web-based interface to calculate phonotactic probability for words and nonwords in English. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 481–487.

Wagner, P. (2005). Great expectations - introspective vs. perceptual prominence ratings and their acoustic correlates. *Interspeech 2005*, 2381–2384.

Wang, L., Bastiaansen, M., Yang, Y., & Hagoort, P. (2011). The influence of information structure on the depth of semantic processing: How focus and pitch accent determine the size of the N400 effect. *Neuropsychologia*, *49*(5), 813–820.

Wang, T., Liu, J., Lee, Y.-h., & Lee, Y.-c. (2020). The interaction between tone and prosodic focus in Mandarin Chinese. *Language and Linguistics*, *21*(2), 331–350.

Warren, P. (1999). Prosody and Language Processing. In S. Garrod & M. Pickering (Eds.), *Language Process* (pp. 155–184). Pychology Press.

Warren, R. M. (1970). Perceptual Restoration of Missing Speech Sounds. *Science*, *167*(3917), 392–393.

Watson, D. G. (2010). The Many Roads to Prominence. In *Psychology of Learning and Motivation* (pp. 163–183, Vol. 52). Elsevier.

Watson, D. G., Tanenhaus, M. K., & Gunlogson, C. A. (2008). Interpreting Pitch Accents in Online Comprehension: H* vs. L+H*. *Cognitive Science*, *32*(7), 1232–1244.

Wayne S., M. (2000). Interaction versus autonomy: A close shave. *Behavioral and Brain Sciences*, *23*(3), 341–342.

Weber, A., & Cutler, A. (2004). Lexical competition in non-native spoken-word recognition. *Journal of Memory and Language*, *50*(1), 1–25.

Weber, A., & Cutler, A. (2006). First-language phonotactics in second-language listening. *The Journal of the Acoustical Society of America*, *119*(1), 597–607.

Weber, A., Grice, M., & Crocker, M. W. (2006). The role of prosody in the interpretation of structural ambiguities: A study of anticipatory eye movements. *Cognition*, *99*(2), B63–B72.

Weber, A., & Scharenborg, O. (2012). Models of spoken-word recognition: Models of word recognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *3*(3), 387–401.

Wells, J. (2006). Chapter 3 Tonicity: Where does the nucleus go? In *English Intonation: An Introduction* (pp. 109–111). University of Cambridge Press.

Wenrich, K. A., Davidson, L. S., & Uchanski, R. M. (2017). Segmental and Suprasegmental Perception in Children Using Hearing Aids. *Journal of the American Academy of Audiology*, *28*(10), 901–912.

Werker, J. F., & Logan, J. S. (1985). Cross-language evidence for three factors in speech perception. *Perception & Psychophysics*, *37*(1), 35–44.

Whalen, D. H., & Liberman, A. M. (1987). Speech perception takes precedence over nonspeech perception. *Science*, *237(4811)*, 168–171.

White, L. (2011). Second language acquisition at the interfaces. *Lingua*, *121*(4), 577–590.

Wood, S. N. (2003). Thin Plate Regression Splines. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *65*(1), 95–114.

Wu, Z., & Ortega-Llebaria, M. (2017). Pitch shape modulates the time course of tone vs pitch-accent identification in Mandarin Chinese. *The Journal of the Acoustical Society of America*, *141*(3), 2263–2276.

Xu, Y. (1999). Effects of tone and focus on the formation and alignment of f0contours. *Journal of Phonetics*, *27*(1), 55–105.

Xue, X., & Lee, J.-K. (2014). The effects of non-native listeners' L1 background and L2 proficiency on the perception of foreign accent for pitch-manipulated native and Chinese accented English speech. *Linguistic Research*, *31*(2), 275–303.

Yang, A., & Chen, A. (2018). The developmental path to adult-like prosodic focus-marking in Mandarin Chinese-speaking children. *First Language*, *38*(1), 26–46.

Ying, H. G. (1996). Multiple Constraints on Processing Ambiguous Sentences: Evidence From Adult L2 Learners. *Language Learning*, *46*(4), 681–711.

Yip, M. C. W. (2017). Probabilistic Phonotactics as a Cue for Recognizing Spoken Cantonese Words in Speech. *Journal of Psycholinguistic Research*, *46*(1), 201–210.

Zhang, J., Duanmu, S., & Chen, Y. (2020). China and Siberia. In C. Gussenhoven & A. Chen (Eds.), *The Oxford Handbook of Language Prosody* (pp. 331–343). Oxford University Press.

Appendix A

Complementary analysis for Chapter 4

This appendix provides a complementary analysis of data in Experiment 1 and 2. The RTs were measured from the start of the stimulus.

A.1 Experiment 1

For Experiment 1, the complementary analysis revealed no main effects or interactions when the data were aggregated by both item and subject. This suggests low item variability in the English pseudo-words. In contrast to the analysis in Section 4.2, the complementary analysis also showed no main effects or interactions when aggregated by subject. The period of initial onset times (prior to the target offset) differed by a few milliseconds across stimuli and conditions (either longer or shorter) before listeners actually heard the stimuli. These small timing differences could result in negligible changes in listeners' RTs under the influence of the artificial accent and prior knowledge.

A.1.1 Masking effect for English listeners

By-item analysis

Effect	DFn	DFd	F	p
pk	1	16	0.16	0.69
modality	1	16	0.02	0.89
accent	1	16	0.97	0.33
pk:modality	1	16	0.57	0.46
pk:accent	1	16	0.003	0.95
modality:accent	1	16	0.02	0.89
pk:modality:accent	1	16	0.20	0.66

Table A.1: Complementary by-item analysis for Experiment 1. English pseudo-word stimuli. ANOVA of masking effect for English listeners. Prior knowledge is referred to as *pk*.

By-subject analysis

Effect	DFn	DFd	F	p
pk	1	76	2.19	0.14
modality	1	76	0.24	0.63
accent	1	76	0.40	0.53
pk:modality	1	76	3.51	0.06
pk:accent	1	76	0.16	0.69
modality:accent	1	76	0.06	0.80
pk:modality:accent	1	76	0.65	0.42

Table A.2: Complementary by-subject analysis for Experiment 1. English pseudo-word stimuli. ANOVA of masking effect for English listeners. Prior knowledge is referred to as *pk*.

A.1.2 Masking effect of Mandarin listeners

By-item analysis

Effect	DFn	DFd	F	p
pk	1	16	0.50	0.49
modality	1	16	0.06	0.81
accent	1	16	0.67	0.43
pk:modality	1	16	0.03	0.86
pk:accent	1	16	0.43	0.52
modality:accent	1	16	0.002	0.97
pk:modality:accent	1	16	0.009	0.93

Table A.3: Complementary by-item analysis for Experiment 1. English pseudo-word stimuli. ANOVA of masking effect for Mandarin listeners. Prior knowledge is referred to as *pk*.

By-subject analysis

Effect	DFn	DFd	F	p
pk	1	76	4.57	0.04
modality	1	76	0.002	0.96
accent	1	76	0.006	0.94
pk:modality	1	76	0.05	0.82
pk:accent	1	76	2.44	0.12
modality:accent	1	76	0.16	0.69
pk:modality:accent	1	76	0.07	0.79

Table A.4: Complementary by-subject analysis for Experiment 1. English pseudo-word stimuli. ANOVA of masking effect for Mandarin listeners. Prior knowledge is referred to as *pk*.

A.2 Experiment 2

The complementary analysis for Experiment 2 revealed no main effects or interactions when the data were aggregated by item and by subject for Mandarin listeners. For English listeners, in the by-item analysis, Accent interacted with Modality (F(1, 16) = 7.52, $p < 0.05$). The post-hoc paired t-test (with Accent as a within-item variable) showed that Accent had a significant effect in the online condition (t(9) = -2.38, $p < 0.05$). Similarly, in the by-subject analysis, Accent also interacted with Modality (F(1, 52) = 6.28, $p < 0.05$). The post-hoc paired t-test showed that Accent had a significant effect in the online condition (t(9) = -3.01, $p < 0.01$).

As shown in Figures A.1 and A.2, the overall masking effect decreased in the online condition with an artificial accent. As depicted in Figure A.3, under the online condition with an artificial accent, RTs to syllables increased more than RTs to phonemes, resulting in a decrease in the masking effect.

A.2.1 Masking effect of Mandarin listeners

By-item analysis

Effect	DFn	DFd	F	p
pk	1	16	0.28	0.60
modality	1	16	0.22	0.64
accent	1	16	0.01	0.97
pk:modality	1	16	0.15	0.70
pk:accent	1	16	0.07	0.79
modality:accent	1	16	1.29	0.27
pk:modality:accent	1	16	0.06	0.81

Table A.5: Complementary by-item analysis for Experiment 2. Mandarin pseudo-word stimuli. ANOVA of masking effect for Mandarin listeners. Prior knowledge is referred to as *pk*.

By-subject analysis

Effect	DFn	DFd	F	p
pk	1	76	0.23	0.63
modality	1	76	0.02	0.90
accent	1	76	0.85	0.36
pk:modality	1	76	0.08	0.78
pk:accent	1	76	0.43	0.51
modality:accent	1	76	0.51	0.48
pk:modality:accent	1	76	0.04	0.84

Table A.6: Complementary by-subject analysis for Experiment 2. Mandarin pseudo-word stimuli. ANOVA of masking effect for Mandarin listeners. Prior knowledge is referred to as *pk*.

A.2.2 Masking effect of English listeners

By-item analysis

Effect	DFn	DFd	F	p
pk	1	16	0.003	0.95
modality	1	16	0.05	0.82
accent	1	16	4.99	0.04
pk:modality	1	16	1.79	0.20
pk:accent	1	16	2.00	0.18
modality:accent	1	16	7.52	< 0.05
pk:modality:accent	1	16	4.36	0.05

Table A.7: Complementary by-item analysis for Experiment 2. Mandarin pseudo-word stimuli. ANOVA of masking effect for English listeners. Prior knowledge is referred to as *pk*.

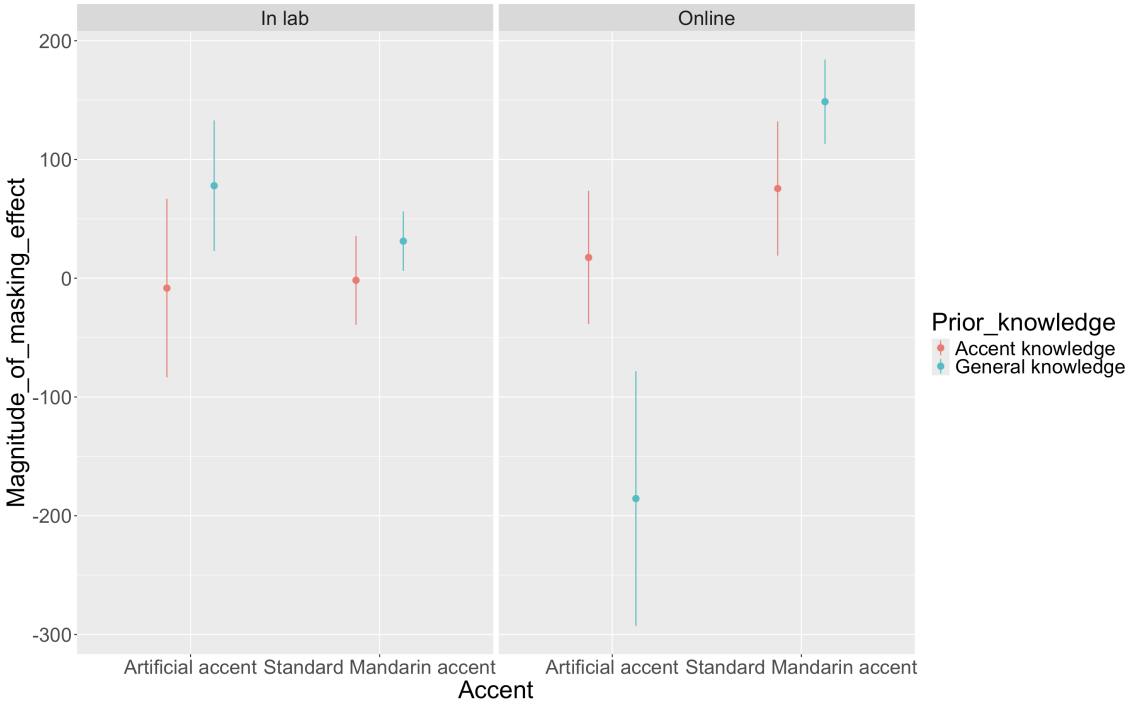


Figure A.1: Complementary by-item analysis for Experiment 2. Mandarin pseudo-word stimuli. Masking effect of English listeners in milliseconds (ms) on y-axis. Conditions of *Accent* (Artificial accent/Standard Mandarin accent) are on x-axis. Fill color indicates the conditions of *Prior knowledge*(Accent knowledge/General knowledge). The facets indicate the conditions of *Modality* (In lab/Online).

By-subject analysis

Effect	DFn	DFd	F	p
pk	1	52	0.01	0.92
modality	1	52	0.09	0.76
accent	1	52	6.10	0.02
pk:modality	1	52	1.46	0.23
pk:accent	1	52	2.26	0.14
modality:accent	1	52	6.28	< 0.05
pk:modality:accent	1	52	3.05	0.09

Table A.8: Complementary by-subject analysis for Experiment 2. Mandarin pseudo-word stimuli. ANOVA of masking effect for English listeners. Prior knowledge is referred to as *pk*.

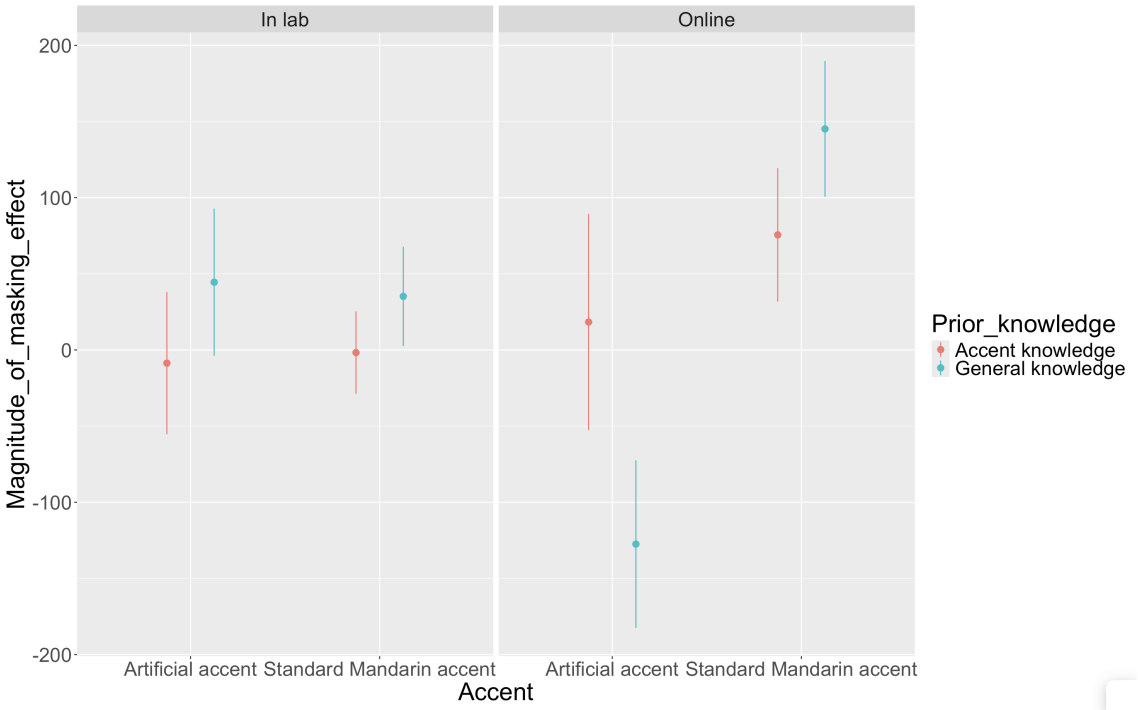


Figure A.2: Complementary by-subject analysis for Experiment 2. Mandarin pseudo-word stimuli. Masking effect of English listeners in milliseconds (ms) on y-axis. Conditions of *Accent* (Artificial accent/Standard Mandarin accent) are on x-axis. Fill color indicates the conditions of *Prior knowledge*(Accent knowledge/General knowledge). The facets indicate the conditions of *Modality* (In lab/Online).

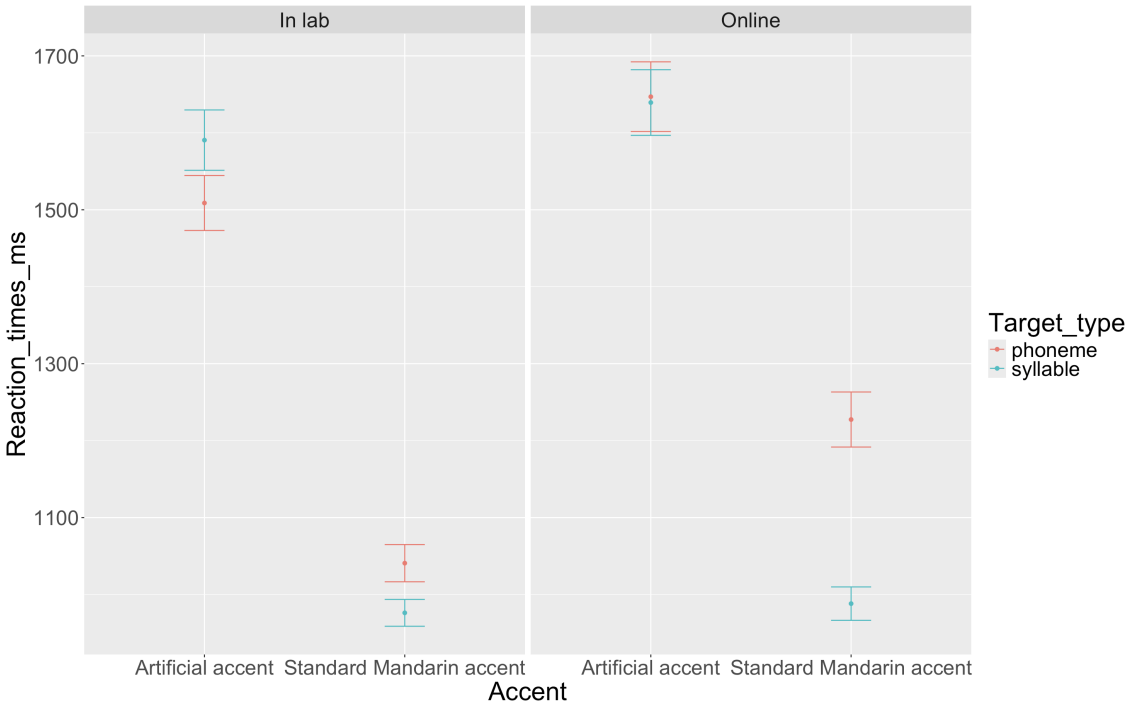


Figure A.3: Complementary analysis for Experiment 2. Mandarin pseudo-word stimuli. RTs to phoneme and syllable targets for English listeners in milliseconds (ms) on y-axis, showing interaction between *Accent* and *Modality*.

Appendix B

Stimuli for Chapter 3

For both English and Mandarin stimuli in *Section A.1* and *A.2*, stimuli in List A and List C are identical, as are List B and List D. However, listeners of List A and B were not provided with prior knowledge on artificial accent, whereas those who completed List C and D were. Phoneme targets were presented in Blocks 1-5 and 11-15, while syllable targets were introduced in Blocks 6-10 and 16-20. The critical targets are in bold.

B.1 English stimuli

B.1.1 Stimuli used for List A and List C (block 1-20)

block1	block2	block3	block4	block5	block6	block7	block8	block9	block10
/d/	/n/	/b/	/k/	/l/	/den/	/ned/	/bɪs/	/kɔm/	/laɪ/
telson	mastet	**basline**	**capnar**	rasto	dagtill	**nedwawn**	**biscri**	**combete**	losken
tengion	**nitnought**	pirthly	**kipsill**	rumcawd	**dentit**	nicknict	balble	curnave	lotler
descate	mitfle	**bilwit**	ganvet	**lantist**	disnawst	nurbike	**bisthawt**	carbeme	**larmest**
tispin	**netleme**	**beshpoy**	**kinchap**	racta	dirchous	neakful	bemple	**compats**	lofpoys
geplo	masto	pedlish	tidloke	rispawt	dabtice	**nedbake**	bensust	**compins**	**larsark**
pilkag	mirtith	mudgewawn	**cusbine**	**louslo**	dadtike	nicknill	**bisfeld**	**compird**	lompon

block11	block12	block13	block14	block15	block16	block17	block18	block19	block20
/ɖʃ/	/ɲ/	/ɦ/	/kʼ/	/ɬ/	/ɖen/	/ɲed/	/ɦɪs/	/kʼɔm/	/ɬaɪ/
tompom	**nethless**	gultish	panco	**lintass**	**dentin**	**nedlert**	backmoot	cursake	lompest
duplic	**nuptar**	dispin	**casman**	rambirt	dagtile	nurtar	**bisfed**	carmeen	**larfo**
dispa	yampus	gungro	purstol	**linty**	**denfle**	**nedbass**	besspoy	**combose**	lubchous
dapmice	yacten	**busthess**	puptish	rimbus	**dentawss**	neeptics	befter	custeen	**larno**
tinlish	**nesstish**	darvo	tambit	**lampin**	dishast	**nedleme**	belden	curbost	**larlish**
dacma	yimbirt	**bilbood**	tanfish	randap	**dentho**	nickdact	**bisred**	karsact	lusquer

B.1.2 Stimuli used for List B and List D (block 1-20)

block1	block2	block3	block4	block5	block6	block7	block8	block9	block10
/d/	/n/	/b/	/k/	/l/	/den/	/ned/	/bɪs/	/kɔm/	/laɪ/
tompom	**nethless**	pultish	ganco	**lintass**	**dentin**	**nedlert**	backmoot	cursake	lompest
duplic	**nuptar**	mispin	**casman**	rambirt	dagtile	nurtar	bisfed	**carmeen**	larfo
dispa	mampus	pungro	gurstol	**linty**	**denfle**	**nedbace**	besspoy	combose	lubchous
dapmice	macten	*busthess*	guptish	**rimbus**	**dentoss**	neeptics	befter	custine	**larno**
tinlish	**nesstish**	marvo	tambit	lampin	dishast	**nedleme**	belden	curbost	**larlish**
dacma	mimbirt	**bilbood**	tanfish	randap	**dentho**	nickdact	bisred	karsact	lusquer

block11	block12	block13	block14	block15	block16	block17	block18	block19	block20
/ʤ/	/ŋ/	/ɓ/	/kʾ/	/ɬ/	/ɖɐn/	/ɲɐd/	/ɓɪs/	/kʾɔm/	/ɬaɯ/
telson	yastet	basline	capnar	rasto	dagtill	**nedwawn**	**biscri**	**combete**	losken
tengion	**nitnot**	girthly	kipsill	**rumcawd**	**dentit**	nicknict	balble	curnave	lotler
descate	yitfle	girthly	**tanvet**	lantist	disnost	nurbike	**bistawt**	carbeme	**larmest**
tispin	**netleme**	beshpoy	kinchap	*racta*	dirchous	neakful	bemple	**compats**	lofpoys
zeplo	yasto	**gedlish**	pidloke	rispawt	dabtice	**nedbake**	bensust	**compins**	**larsark**
zilkag	yirtith	gudgewawn	cusbine	**louslo**	dadtike	nicknill	*bisteld*	**compird**	lompon

B.2 Mandarin stimuli

B.2.1 Stimuli used for List A and List C (block 1-20)

block1	block2	block3	block4	block5	block6	block7	block8	block9	block10
/tʰ/	/n/	/x/	/kʰ/	/l/	/tʰan/	/niŋ/	/xən/	/kʰoŋ/	/lau/
di4 lu3	mao4 zhu1	**hai4 ji4**	**kuan4 dui4**	ruan4 ji2	ti4 bing1	**ning4 ji2**	**hen4 zong1**	**kong4 tai4**	lie4 lin2
dai4 cai4	**nan4 hui4**	gai4 piao1	**kong4 ji2**	ru4 lun4	**tan4 ju2**	nang4 shi2	hai4 li3	kai4 zhong4	long4 ke4
tao4 kuo4	mang4 bai4	**han4 kao4**	gang4 li4	**lan4 qi2**	tu4 cong4	nao4 li2	**hen4 fu1**	ke4 rong2	**lao4 di4**
dao4 hu1	**ning4 lang2**	**hu4 zhang3**	**ku4 yuan2**	ren4 chi4	tie4 yong3	nai4 zong1	hao4 ni4	**kong4 dai4**	lu4 wen1
deng4 fa4	man4 ding4	gan4 zan1	pai4 hua2	rao4 di1	tian4 quan1	**ning4 li2**	hu4 di4	**kong4 li3**	**lao 4 ku1**
dian4 sai4	man4 xiang3	gou4 jue2	**ke4 tu2**	**liang4 gao1**	teng4 jie2	neng4 hao1	**hen4 ling2**	**kong4 tan2**	liu4 xi2

block11	block12	block13	block14	block15	block16	block17	block18	block19	block20
/tʰ/	/ɲ/	/h/	/k'/	/ɬ/	/tʰan/	/ɲiŋ/	/hən/	/k'oŋ/	/ɬəu/
dai4 qi1	**neng4 xi1**	tao4 bu3	peng4 liang4	**lu4 tao1**	**tan4 zheng3**	**ning4 cheng1**	hua4 pian1	kai4 yan4	luo4 pan4
tang4 feng1	**ni4 zhuang1**	te4 huan3	**ke4 huo3**	ran4 biao3	dui4 pan4	nao4 pan4	**hen4 teng2**	kan4 sheng1	**lao4 lun4**
tai4 chang2	yan4 du4	tan4 wen1	peng4 xing1	**lian4 shu4**	**tan4 qian1**	**ning4 fu4**	hu4 pian1	**kong4 fang2**	lu4 ban4
tu4 zhe2	ya4 dou3	**hai4 liang2**	pou4 zi3	ru4 hao2	**tan4 yun4**	nan4 jiang3	hou4 xiao1	ka4 jia1	**lao4 yan2**
deng4 bei4	**nian4 feng1**	tang4 cai2	pu4 lun2	**liao4 gui1**	duo4 ban4	**ning4 gao1**	hong4 dian4	ku4 fu2	**lao4 hang2**
teng4 de2	yan4 deng1	**hui4 jie2**	pan4 shu3	ruo4 ba3	**tan4 hui3**	na4 pu4	**hen4 shi4**	kuo4 de2	li4 su1

B.2.2 Stimuli used for List B and List D (block 1-20)

block1	block2	block3	block4	block5	block6	block7	block8	block9	block10
/tʰ/	/n/	/x/	/kʰ/	/l/	/tʰan/	/niŋ/	/xən/	/kʰoŋ/	/lau/
dai4 qi1	**neng4 xi1**	gao4 bu3	peng4 liang4	**lu4 tao1**	**tan4 zheng3**	**ning4 cheng1**	hua4 pian1	kai4 yan4	luo4 pan4
tang4 feng1	**ni4 zhuang1**	ge4 huan3	**ke4 huo3**	ran4 biao3	dui4 pan4	nao4 pan4	**hen4 teng2**	kan4 sheng1	**lao4 lun4**
tai4 chang2	mang4 du4	gan4 wen1	peng4 xing1	**lian4 shu4**	**tan4 qian1**	**ning4 fu4**	hu4 pian1	**kong4 fang2**	lu4 ban4
tu4 zhe2	ma4 dou3	**hai4 liang2**	pou4 zi3	ru4 hao2	**tan4 yun4**	nan4 jiang3	hou4 xiao1	ka4 jia1	**lao4 yan2**
deng4 bei4	**nian4 feng1**	gang4 cai2	pu4 lun2	**liao4 gui1**	duo4 ban4	**ning4 gao1**	hong4 dian4	ku4 fu2	**lao4 hang2**
vteng4 de2	man4 deng1	**hui4 jie2**	pan4 shu3	ruo4 ba3	**tan4 hui3**	na4 pu4	**hen4 shi4**	kuo4 de2	li4 su1

block11	block12	block13	block14	block15	block16	block17	block18	block19	block20
/tʰ/	/ɲ/	/h/	/k'/	/ɬ/	/tʰan/	/ɲiŋ/	/hən/	/k'oŋ/	/ɬəu/
di4 lu3	yao4 zhu1	**hai4 ji4**	**kuan4 dui4**	ruan4 ji2	ti4 bing1	**ning4 ji2**	**hen4 zong1**	**kong4 tai4**	lie4 lin2
dai4 cai4	**nan4 hui4**	tai4 piao1	**kong4 ji2**	ru4 lun4	**tan4 ju2**	nang4 shi2	hai4 li3	kai4 zhong4	long4 ke4
tao4 kuo4	yang4 bai4	**han4 kao4**	pang4 li4	**lan4 qi2**	tu4 cong4	nao4 li2	**hen4 fu1**	ke4 rong2	**lao4 di4**
dao4 hu1	**ning4 lang2**	**hu4 zhang3**	**ku4 yuan2**	ren4 chi4	tie4 yong3	nai4 zong1	hao4 ni4	**kong4 dai4**	lu4 wen1
deng4 fa4	yan4 ding4	tan4 zan1	pai4 hua2	rao4 di1	tian4 quan1	**ning4 li2**	hu4 di4	**kong4 li3**	**lao 4 ku1**
dian4 sai4	yan4 xiang3	tou4 jue2	**ke4 tu2**	**liang4 gao1**	teng4 jie2	neng4 hao1	**hen4 ling2**	**kong4 tan2**	liu4 xi2

Appendix C

Stimuli for Chapter 6

C.1 English stimuli

For English stimuli, targets accentuated with contrastive pitch accent are capitalized.

C.1.1 Critical sentences

1. *cabbage-captain*

Here is a cabbage. Susan is going to shred the CABBAGE after school.

Here is a captain. Susan is going to shred the CABBAGE after school.

Here is a cabbage. Susan is going to shred the cabbage after school.

Here is a captain. Susan is going to shred the cabbage after school.

Here is a cabbage. Susan is going to drink the CABBAGE after school.

Here is a captain. Susan is going to drink the CABBAGE after school.

Here is a cabbage. Susan is going to drink the cabbage after school.

Here is a captain. Susan is going to drink the cabbage after school.

2. *candle-candy*

Here is a candle. Aaliyah is going to light the CANDLE before supper.

Here is a candy. Aaliyah is going to light the CANDLE before supper.

Here is a candle. Aaliyah is going to light the candle before supper.

Here is a candy. Aaliyah is going to light the candle before supper.

Here is a candle. Aaliyah is going to pour the CANDLE before supper.

Here is a candy. Aaliyah is going to pour the CANDLE before supper.

Here is a candle. Aaliyah is going to pour the candle before supper.

Here is a candy. Aaliyah is going to pour the candle before supper.

3.*pigeon-pitcher*

Here is a pigeon. Susan is going to feed the PIGEON after work.

Here is a pitcher. Susan is going to feed the PIGEON after work.

Here is a pigeon. Susan is going to feed the pigeon after work.

Here is a pitcher. Susan is going to feed the pigeon after work.

Here is a pigeon. Susan is going to wear the PIGEON after work.

Here is a pitcher. Susan is going to wear the PIGEON after work.

Here is a pigeon. Susan is going to wear the pigeon after work.

Here is a pitcher. Susan is going to wear the pigeon after work.

4. *walnut-wallet*

Here is a walnut. Regis is going to crack the WALNUT at dinner.

Here is a wallet. Regis is going to crack the WALNUT at dinner.

Here is a walnut. Regis is going to crack the walnut at dinner.

Here is a wallet. Regis is going to crack the walnut at dinner.

Here is a walnut. Regis is going to grease the WALNUT at dinner.

Here is a wallet. Regis is going to grease the WALNUT at dinner.

Here is a walnut. Regis is going to grease the walnut at dinner.

Here is a wallet. Regis is going to grease the walnut at dinner.

5.*pencil-pendent*

Here is a pencil. Emilia is going to sharpen the PENCIL during the painting class.

Here is a pendent. Emilia is going to sharpen the PENCIL during the painting class.

Here is a pencil. Emilia is going to sharpen the pencil during the painting class.

Here is a pendent. Emilia is going to sharpen the pencil during the painting class.

Here is a pencil. Emilia is going to sprinkle the PENCIL during the painting class.

Here is a pendent. Emilia is going to sprinkle the PENCIL during the painting class.

Here is a pencil. Emilia is going to sprinkle the pencil during the painting class.

Here is a pendent. Emilia is going to sprinkle the pencil during the painting class.

6. *hamster-hammock*

Here is a hamster. Lariya is going to feed the HAMSTER before dinner.

Here is a hammock. Lariya is going to feed the HAMSTER before dinner.

Here is a hamster. Lariya is going to feed the hamster before dinner.

Here is a hammock. Lariya is going to feed the hamster before dinner.

Here is a hamster. Lariya is going to melt the HAMSTER before dinner.

Here is a hammock. Lariya is going to melt the HAMSTER before dinner.

Here is a hamster. Lariya is going to melt the hamster before dinner.

Here is a hammock. Lariya is going to melt the hamster before dinner.

7. *banjo-bandage*

Here is a banjo. Bryden is going to tune the BANJO before music class.

Here is a bandage. Bryden is going to tune the BANJO before music class.

Here is a banjo. Bryden is going to tune the banjo before music class.

Here is a bandage. Bryden is going to tune the banjo before music class.

Here is a banjo. Bryden is going to unlock the BANJO before music class.

Here is a bandage. Bryden is going to unlock the BANJO before music class.

Here is a banjo. Bryden is going to unlock the banjo before music class.

Here is a bandage. Bryden is going to unlock the banjo before music class.

8. *shellfish-shelving*

Here is a shellfish. Grandin is going to eat the SHELLFISH during lunchtime.

Here is a shelving. Grandin is going to eat the SHELLFISH during lunchtime.

Here is a shellfish. Grandin is going to eat the shellfish during lunchtime.

Here is a shelving. Grandin is going to eat the shellfish during lunchtime.

Here is a shellfish. Grandin is going to teach the SHELLFISH during lunchtime.

Here is a shelving. Grandin is going to teach the SHELLFISH during lunchtime.

Here is a shellfish. Grandin is going to teach the shellfish during lunchtime.

Here is a shelving. Grandin is going to teach the shellfish during lunchtime.

9.*tuba-tulip*

Here is a tuba. Emma is going to play the TUBA prior to the music exam.

Here is a tulip. Emma is going to play the TUBA prior to the music exam.

Here is a tuba. Emma is going to play the tuba prior to the music exam.

Here is a tulip. Emma is going to play the tuba prior to the music exam.

Here is a tuba. Emma is going to drain the TUBA prior to the music exam.

Here is a tulip. Emma is going to drain the TUBA prior to the music exam.

Here is a tuba. Emma is going to drain the tuba prior to the music exam.

Here is a tulip. Emma is going to drain the tuba prior to the music exam.

10. *turnip-turkey*

Here is a turnip. Benjamin is going to peel the TURNIP at teatime.

Here is a turkey. Benjamin is going to peel the TURNIP at teatime.

Here is a turnip. Benjamin is going to peel the turnip at teatime.

Here is a turkey. Benjamin is going to peel the turnip at teatime.

Here is a turnip. Benjamin is going to mock the TURNIP at teatime.

Here is a turkey. Benjamin is going to mock the TURNIP at teatime.

Here is a turnip. Benjamin is going to mock the turnip at teatime.

Here is a turkey. Benjamin is going to mock the turnip at teatime.

11.*jumper-jungle*

Here is a jumper. Samantha is going to wear the JUMPER during the winter.

Here is a jungle. Samantha is going to wear the JUMPER during the winter.

Here is a jumper. Samatha is going to wear the jumper during the winter.

Here is a jungle. Samatha is going to wear the jumper during the winter.

Here is a jumper. Samantha is going to grind the JUMPER during the winter.

Here is a jungle. Samantha is going to grind the JUMPER during the winter.

Here is a jumper. Samatha is going to grind the jumper during the winter.

Here is a jungle. Samatha is going to grind the jumper during the winter.

12. *pillow-pillar*

Here is a pillow. Samuel is going to smooth the PILLOW before bedtime.

Here is a pillar. Samuel is going to smooth the PILLOW before bedtime.

Here is a pillow. Samuel is going to smooth the pillow before bedtime.

Here is a pillar. Samuel is going to smooth the pillow before bedtime.

Here is a pillow. Samuel is going to inject the PILLOW before bedtime.

Here is a pillar. Samuel is going to inject the PILLOW before bedtime.

Here is a pillow. Samuel is going to inject the pillow before bedtime.

Here is a pillar. Samuel is going to inject the pillow before bedtime.

13. *carpet-cartoon*

Here is a carpet. Catherine is going to vacuum the CARPET before work.

Here is a cartoon. Catherine is going to vacuum the CARPET before work.

Here is a carpet. Catherine is going to vacuum the carpet before work.

Here is a cartoon. Catherine is going to vacuum the carpet before work.

Here is a carpet. Catherine is going to poison the CARPET before work.

Here is a cartoon. Catherine is going to poison the CARPET before work.

Here is a carpet. Catherine is going to poison the carpet before work.

Here is a cartoon. Catherine is going to poison the carpet before work.

14.*kitten-kitchen*

Here is a kitten. Gregory is going to groom the KITTEN in her spare time.

Here is a kitchen. Gregory is going to groom the KITTEN in her spare time.

Here is a kitten. Gregory is going to groom the kitten in her spare time.

Here is a kitchen. Gregory is going to groom the kitten in her spare time.

Here is a kitten. Gregory is going to drive the KITTEN in her spare time.

Here is a kitchen. Gregory is going to drive the KITTEN in her spare time.

Here is a kitten. Gregory is going to drive the kitten in her spare time.

Here is a kitchen. Gregory is going to drive the kitten in her spare time.

15. *plaster-platter*

Here is a plaster. Cristine is going to stick the PLASTER before taking a shower.

Here is a platter. Cristine is going to stick the PLASTER before taking a shower.

Here is a plaster. Cristine is going to stick the plaster before taking a shower.

Here is a platter. Cristine is going to stick the plaster before taking a shower.

Here is a plaster. Cristine is going to fight the PLASTER before taking a shower.

Here is a platter. Cristine is going to fight the PLASTER before taking a shower.

Here is a plaster. Cristine is going to fight the plaster before taking a shower.

Here is a platter. Cristine is going to fight the plaster before taking a shower.

16. *barbell-barber*

Here is a barbell. Frank is going to lift the BARBELL before breakfast.

Here is a barber. Frank is going to lift the BARBELL before breakfast.

Here is a barbell. Frank is going to lift the barbell before breakfast.

Here is a barber. Frank is going to lift the barbell before breakfast.

Here is a barbell. Frank is going to swallow the BARBELL before breakfast.

Here is a barber. Frank is going to swallow the BARBELL before breakfast.

Here is a barbell. Frank is going to swallow the barbell before breakfast.

Here is a barber. Frank is going to swallow the barbell before breakfast.

17. *trouser-trowel*

Here is a trousers. Dapper is going to fold the TROUSERS before going to sleep.

Here is a trowel. Dapper is going to fold the TROUSERS before going to sleep.

Here is a trousers. Dapper is going to fold the trousers before going to sleep.

Here is a trowel. Dapper is going to fold the trousers before going to sleep.

Here is a trousers. Dapper is going to punch the TROUSERS before going to sleep.

Here is a trowel. Dapper is going to punch the TROUSERS before going to sleep.

Here is a trousers. Dapper is going to punch the trousers before going to sleep.

Here is a trowel. Dapper is going to punch the trousers before going to sleep.

18. *crocus-crowbar*

Here is a crocus. Alex is going to pick the CROCUS before class.

Here is a crowbar. Alex is going to pick the CROCUS before class.

Here is a crocus. Alex is going to pick the crocus before class.

Here is a crowbar. Alex is going to pick the crocus before class.

Here is a crocus. Alex is going to build the CROCUS before class.

Here is a crowbar. Alex is going to build the CROCUS before class.

Here is a crocus. Alex is going to build the crocus before class.

Here is a crowbar. Alex is going to build the crocus before class.

19. *bacon-baker*

Here is a bacon. Rachel is going to fry the BACON in the morning.

Here is a baker. Rachel is going to fry the BACON in the morning.

Here is a bacon. Rachel is going to fry the bacon in the morning.

Here is a baker. Rachel is going to fry the bacon in the morning.

Here is a bacon. Rachel is going to soften the BACON in the morning.

Here is a baker. Rachel is going to soften the BACON in the morning.

Here is a bacon. Rachel is going to soften the bacon in the morning.

Here is a baker. Rachel is going to soften the bacon in the morning.

20. *slingshot-slinky*

Here is a slingshot. Raymon is going to shoot the SLINGSHOT while watching TV.

Here is a slinky. Raymon is going to shoot the SLINGSHOT while watching TV.

Here is a slingshot. Raymon is going to shoot the slingshot while watching TV.

Here is a slinky. Raymon is going to shoot the slingshot while watching TV.

Here is a slingshot. Raymon is going to write the SLINGSHOT while watching TV.

Here is a slinky. Raymon is going to write the SLINGSHOT while watching TV.

Here is a slingshot. Raymon is going to write the slingshot while watching TV.

Here is a slinky. Raymon is going to write the slingshot while watching TV.

21. *keyboard-keyring*

Here is a keyboard. Katie is going to play the KEYBOARD after music class.

Here is a keyring. Katie is going to play the KEYBOARD after music class.

Here is a keyboard. Katie is going to play the keyboard after music class.

Here is a keyring. Katie is going to play the keyboard after music class.

Here is a keyboard. Katie is going to scold the KEYBOARD after music class.

Here is a keyring. Katie is going to scold the KEYBOARD after music class.

Here is a keyboard. Katie is going to scold the keyboard after music class.

Here is a keyring. Katie is going to scold the keyboard after music class.

22. *whisker-wizard*

Here is a whisker. Patrick is going to trim the WHISKER during her lunch hour.

Here is a wizard. Patrick is going to trim the WHISKER during her lunch hour.

Here is a whisker. Patrick is going to trim the whisker during her lunch hour.

Here is a wizard. Patrick is going to trim the whisker during her lunch hour.

Here is a whisker. Patrick is going to unload the WHISKER during her lunch hour.

Here is a wizard. Patrick is going to unload the WHISKER during her lunch hour.

Here is a whisker. Patrick is going to unload the whisker during her lunch hour.

Here is a wizard. Patrick is going to unload the whisker during her lunch hour.

23. *sandal-sandwich*

Here is a sandal. Caroline is going to buckle the SANDAL before hanging out.

Here is a sandwich. Caroline is going to buckle the SANDAL before hanging out.

Caroline is a sandal. Susan is going to buckle the sandal before hanging out.

Caroline is a sandwich. Susan is going to buckle the sandal before hanging out.

Here is a sandal. Caroline is going to pierce the SANDAL before hanging out.

Here is a sandwich. Caroline is going to pierce the SANDAL before hanging out.

Caroline is a sandal. Susan is going to pierce the sandal before hanging out.

Caroline is a sandwich. Susan is going to pierce the sandal before hanging out.

24. *lifeboat-lifeguard*

Here is a lifeboat. Jack is going to row the LIFEBOAT on weekends.

Here is a lifeguard. Jack is going to row the LIFEBOAT on weekends.

Here is a lifeboat. Jack is going to row the lifeboat on weekends.

Here is a lifeguard. Jack is going to row the lifeboat on weekends.

Here is a lifeboat. Jack is going to twist the LIFEBOAT on weekends.

Here is a lifeguard. Jack is going to twist the LIFEBOAT on weekends.

Here is a lifeboat. Jack is going to twist the lifeboat on weekends.

Here is a lifeguard. Jack is going to twist the lifeboat on weekends.

25. *walnut-wallet*

Here is a lady. Jenny is going to greet the LADY during the party.

Here is a ladle. Jenny is going to greet the LADY during the party.

Here is a lady. Jenny is going to greet the lady during the party

Here is a ladle. Jenny is going to greet the lady during the party.

Here is a lady. Jenny is going to discover the LADY during the party.

Here is a ladle. Jenny is going to discover the LADY during the party.

Here is a lady. Jenny is going to discover the lady during the party.

Here is a ladle. Jenny is going to discover the lady during the party.

26. *badger-batter*

Here is a badger. Dennis is going to hunt the BADGER tomorrow.

Here is a batter. Dennis is going to hunt the BADGER tomorrow.

Here is a badger. Dennis is going to hunt the badger tomorrow.

Here is a batter. Dennis is going to hunt the badger tomorrow.

Here is a badger. Dennis is going to smear the BADGER tomorrow.

Here is a batter. Dennis is going to smear the BADGER tomorrow.

Here is a badger. Dennis is going to smear the badger tomorrow.

Here is a batter. Dennis is going to smear the badger tomorrow.

27. *ketchup-kettle*

Here is a ketchup. Ruth is going to squirt the KETCHUP at brunch.

Here is a kettle. Ruth is going to squirt the KETCHUP at brunch.

Here is a ketchup. Ruth is going to squirt the ketchup at brunch.

Here is a kettle. Ruth is going to squirt the ketchup at brunch.

Here is a ketchup. Ruth is going to nip the KETCHUP at brunch.

Here is a kettle. Ruth is going to nip the KETCHUP at brunch.

Here is a ketchup. Ruth is going to nip the ketchup at brunch.

Here is a kettle. Ruth is going to nip the ketchup at brunch.

28. *collar-column*

Here is a collar. Jered is going to button the COLLAR before dating.

Here is a column. Jered is going to button the COLLAR before dating.

Here is a collar. Jered is going to button the collar before dating.

Here is a column. Jered is going to button the collar before dating.

Here is a collar. Jered is going to grill the COLLAR before dating.

Here is a column. Jered is going to grill the COLLAR before dating.

Here is a collar. Jered is going to grill the collar before dating.

Here is a column. Jered is going to grill the collar before dating.

29. *butter-button*

Here is a butter. Maria is going to spread the BUTTER before cooking.

Here is a button. Maria is going to spread the BUTTER before cooking.

Here is a butter. Maria is going to spread the butter before cooking.

Here is a button. Maria is going to spread the butter before cooking.

Here is a button. Maria is going to drag the BUTTER before cooking.

Here is a butter. Maria is going to drag the BUTTER before cooking.

Here is a button. Maria is going to drag the butter before cooking.

Here is a butter. Maria is going to drag the butter before cooking.

30. *panda-pansy*

Here is a panda. Taylor is going to cuddle the PANDA during working.

Here is a pansy. Taylor is going to cuddle the PANDA during working.

Here is a panda. Taylor is going to cuddle the panda during working.

Here is a pansy. Taylor is going to cuddle the panda during working.

Here is a panda. Taylor is going to roll the PANDA during working.

Here is a pansy. Taylor is going to roll the PANDA during working.

Here is a panda. Taylor is going to roll the panda during working.

Here is a pansy. Taylor is going to roll the panda during working.

31. *radish-rattle*

Here is a radish. Heser is going to slice the RADISH while listening to music.

Here is a rattle. Heser is going to slice the RADISH while listening to music.

Here is a radish. Heser is going to slice the radish while listening to music.

Here is a rattle. Heser is going to slice the radish while listening to music.

Here is a radish. Heser is going to embrace the RADISH while listening to music.

Here is a rattle. Heser is going to embrace the RADISH while listening to music.

Here is a radish. Heser is going to embrace the radish while listening to music.

Here is a rattle. Heser is going to embrace the radish while listening to music.

32. *hamper-hammer*

Here is a hamper. Aaron is going to unpack the HAMPER before picnicking.

Here is a hammer. Aaron is going to unpack the HAMPER before picnicking.

Here is a hamper. Aaron is going to unpack the hamper before picnicking.

Here is a hammer. Aaron is going to unpack the hamper before picnicking.

Here is a hamper. Aaron is going to sing the HAMPER before picnicking.

Here is a hammer. Aaron is going to sing the HAMPER before picnicking.

Here is a hamper. Aaron is going to sing the hamper before picnicking.

Here is a hammer. Aaron is going to sing the hamper before picnicking.

33. *beaker-beetle*

Here is a beaker. Diana is going to wash the BEAKER before chemistry class.

Here is a beetle. Diana is going to wash the BEAKER before chemistry class.

Here is a beaker. Diana is going to wash the beaker before chemistry class.

Here is a beetle. Diana is going to wash the beaker before chemistry class.

Here is a beaker. Diana is going to guzzle the BEAKER before chemistry class.

Here is a beetle. Diana is going to guzzle the BEAKER before chemistry class.

Here is a beaker. Diana is going to guzzle the beaker before chemistry class.

Here is a beetle. Diana is going to guzzle the beaker before chemistry class.

34. *sapling-saddle*

Here is a sapling. Hossein is going to plant the SAPLING in this afternoon.

Here is a saddle. Hossein is going to plant the SAPLING in this afternoon.

Here is a sapling. Hossein is going to plant the sapling in this afternoon.

Here is a saddle. Hossein is going to plant the sapling in this afternoon.

Here is a sapling. Hossein is going to murder the SAPLING in this afternoon.

Here is a saddle. Hossein is going to murder the SAPLING in this afternoon.

Here is a sapling. Hossein is going to murder the sapling in this afternoon.

Here is a saddle. Hossein is going to murder the sapling in this afternoon.

35. *lightbulb-lighthouse*

Here is a lightbulb. Virginia is going to change the LIGHTBULB in a week.

Here is a lighthouse. Virginia is going to change the LIGHTBULB in a week.

Here is a lightbulb. Virginia is going to change the lightbulb in a week.

Here is a lighthouse. Virginia is going to change the lightbulb in a week.

Here is a lightbulb. Virginia is going to define the LIGHTBULB in a week.

Here is a lighthouse. Virginia is going to define the LIGHTBULB in a week.

Here is a lightbulb. Virginia is going to define the lightbulb in a week.

Here is a lighthouse. Virginia is going to define the lightbulb in a week.

36. *lipstick-liquor*

Here is a lipstick. Julia is going to apply the LIPSTICK while answering the phone call.

Here is a liquor. Julia is going to apply the LIPSTICK while answering the phone call.

Here is a lipstick. Julia is going to apply the lipstick while answering the phone call.

Here is a liquor. Julia is going to apply the lipstick while answering the phone call.

Here is a lipstick. Julia is going to chase the LIPSTICK while answering the phone call.

Here is a liquor. Julia is going to chase the LIPSTICK while answering the phone call.

Here is a lipstick. Julia is going to chase the lipstick while answering the phone call.

Here is a liquor. Julia is going to chase the lipstick while answering the phone call.

37. *litter-liver*

Here is a litter. Adam is going to pickup the LITTER next Tuesday.

Here is a liver. Adam is going to pickup the LITTER next Tuesday.

Here is a litter. Adam is going to pickup the litter next Tuesday.

Here is a liver. Adam is going to pickup the litter next Tuesday.

Here is a litter. Adam is going to climb the LITTER next Tuesday.

Here is a liver. Adam is going to climb the LITTER next Tuesday.

Here is a litter. Adam is going to climb the litter next Tuesday.

Here is a liver. Adam is going to climb the litter next Tuesday.

38.*lily-lizard*

Here is a lily. Adam is going to water the LILY before going on holiday.

Here is a lizard. Adam is going to water the LILY before going on holiday.

Here is a lily. Adam is going to water the lily before going on holiday.

Here is a lizard. Adam is going to water the lily before going on holiday.

Here is a lily. Adam is going to furnish the LILY before going on holiday.

Here is a lizard. Adam is going to furnish the LILY before going on holiday.

Here is a lily. Adam is going to furnish the lily before going on holiday.

Here is a lizard. Adam is going to furnish the lily before going on holiday

39.*corset-corkscrew*

Here is a corset. Josey is going to laceup the CORSET before the party.

Here is a corkscrew. Josey is going to laceup the CORSET before the party.

Here is a corset. Josey is going to laceup the corset before the party.

Here is a corkscrew. Josey is going to laceup the corset before the party.

Here is a corset. Josey is going to mash the CORSET before the party.

Here is a corkscrew. Josey is going to mash the CORSET before the party.

Here is a corset. Josey is going to mash the corset before the party.

Here is a crokscrew. Josey is going to mash the croset before the party.

40.*lion-lilac*

Here is a lion. Neeson is going to tame the LION in two weeks

Here is a lilac. Neeson is going to tame the LION in two weeks

Here is a lion. Neeson is going to tame the lion in two weeks

Here is a lilac. Neeson is going to tame the lion in two weeks

Here is a lion. Neeson is going to wind the LION in two weeks

Here is a lilac. Neeson is going to wind the LION in two weeks

Here is a lion. Neeson is going to wind the lion in two weeks

Here is a lilac. Neeson is going to wind the lion in two weeks

C.1.2 Filler sentences

1.*keyring-slingshot*

Here is a keyring. Victoria is going to jingle the KEYRING while driving.

Here is a slingshot. Victoria is going to jingle the KEYRING while driving.

Here is a keyring. Victoria is going to jingle the keyring while driving.

Here is a slingshot. Victoria is going to jingle the keyring while driving.

Here is a keyring. Victoria is going to soak the KEYRING while driving.

Here is a slingshot. Victoria is going to soak the KEYRING while driving.

Here is a keyring. Victoria is going to soak the keyring while driving.

Here is a slingshot. Victoria is going to soak the keyring while driving.

2. *corkscrew-bacon*

Here is a corkscrew. Douglas is going to pick up the CORKSCREW on the ground.

Here is a bacon. Douglas is going to pick up the CORKSCREW on the ground.

Here is a corkscrew. Douglas is going to pick up the corkscrew on the ground.

Here is a bacon. Douglas is going to pick up the corkscrew on the ground.

Here is a corkscrew. Douglas is going to drive the CORKSCREW on the ground.

Here is a bacon. Douglas is going to drive the CORKSCREW on the ground.

Here is a corkscrew. Douglas is going to drive the corkscrew on the ground.

Here is a bacon. Douglas is going to drive the corkscrew on the ground.

3.*slinky-candle*

Here is a slinky. Susan is going to play with the SLINKY during the performance.

Here is a candle. Susan is going to play with the SLINKY during the performance.

Here is a slinky. Susan is going to play with the slinky during the performance.

Here is a candle. Susan is going to play with the slinky during the performance.

Here is a slinky. Susan is going to wear the SLINKY during the performance.

Here is a candle. Susan is going to wear the SLINKY during the performance.

Here is a slinky. Susan is going to wear the slinky during the performance.

Here is a candle. Susan is going to wear the slinky during the performance.

4. *sandwich-cabbage*

Here is a sandwich. Zack is going to heat the SANDWICH before breakfast.

Here is a cabbage. Zack is going to heat the SANDWICH before breakfast.

Here is a sandwich. Zack is going to heat the sandwich before breakfast.

Here is a cabbage. Zack is going to heat the sandwich before breakfast.

Here is a sandwich. Zack is going to drink the SANDWICH before breakfast.

Here is a cabbage. Zack is going to drink the SANDWICH before breakfast.

Here is a sandwich. Zack is going to drink the sandwich before breakfast.

Here is a cabbage. Zack is going to drink the sandwich before breakfast.

5.*captain-shellfish*

Here is a captain. Kylia is going to talk to the CAPTAIN for an half hour.

Here is a shellfish. Kylia is going to talk to the CAPTAIN for an half hour.

Here is a captain. Kylia is going to talk to the captain for an half hour.

Here is a shellfish. Kylia is going to talk to the captain for an half hour.

Here is a captain. Kylia is going to sprinkle the CAPTAIN for an half hour.

Here is a shellfish. Kylia is going to sprinkle the CAPTAIN for an half hour.

Here is a captain. Kylia is going to sprinkle the captain for an half hour.

Here is a shellfish. Kylia is going to sprinkle the captain for an half hour.

6. *wizard-corset*

Here is a wizard. Peter is going to dress like the WIZARD at the party.

Here is a corset. Peter is going to dress like the WIZARD at the party.

Here is a wizard. Peter is going to dress like the wizard at the party.

Here is a corset. Peter is going to dress like the wizard at the party.

Here is a wizard. Peter is going to pinch the WIZARD at the party.

Here is a corset. Peter is going to pinch the WIZARD at the party.

Here is a wizard. Peter is going to pinch the wizard at the party.

Here is a corset. Peter is going to pinch the wizard at the party.

7. *batter-crocus*

Here is a batter. Christina is going to train the BATTER in her vacation.

Here is a crocus. Christina is going to train the BATTER in her vacation.

Here is a batter. Christina is going to train the batter in her vacation.

Here is a crocus. Christina is going to train the batter in her vacation.

Here is a batter. Christina is going to fly with the BATTER in her vacation.

Here is a crocus. Christina is going to fly with the BATTER in her vacation.

Here is a batter. Christina is going to fly with the batter in her vacation.

Here is a crocus. Christina is going to fly with the batter in her vacation.

8.*kettle-trouser*

Here is a kettle. Kyle is going to fill the KETTLE while eating.

Here is a trouser. Kyle is going to fill the KETTLE while eating.

Here is a kettle. Kyle is going to fill the kettle while eating.

Here is a trouser. Kyle is going to fill the kettle while eating.

Here is a kettle. Kyle is going to walk the KETTLE while eating.

Here is a trouser. Kyle is going to walk the KETTLE while eating.

Here is a kettle. Kyle is going to walk the kettle while eating.

Here is a trouser. Kyle is going to walk the kettle while eating.

9. *column-barbell*

Here is a column. Lauren is going to build the COLUMN in a month.

Here is a barbell. Lauren is going to build the COLUMN in a month.

Here is a column. Lauren is going to build the column in a month.

Here is a barbell. Lauren is going to build the column in a month.

Here is a column. Lauren is going to drain the COLUMN in a month.

Here is a barbell. Lauren is going to drain the COLUMN in a month.

Here is a column. Lauren is going to drain the column in a month.

Here is a barbell. Lauren is going to drain the column in a month.

10. *button-plaster*

Here is a button. Walter is going to fasten with the BUTTON before work.

Here is a plaster. Walter is going to fasten with the BUTTON before work.

Here is a button. Walter is going to fasten with the button before work.

Here is a plaster. Walter is going to fasten with the button before work.

Here is a button. Walter is going to paint the BUTTON before work.

Here is a plaster. Walter is going to paint the BUTTON before work.

Here is a button. Walter is going to paint the button before work.

Here is a plaster. Walter is going to paint the button before work.

11. *pansy-kitten*

Here is a pansy. Joe is going to plant the PANSY during her lunch hour.

Here is a kitten. Joe is going to plant the PANSY during her lunch hour.

Here is a pansy. Joe is going to plant the pansy during her lunch hour.

Here is a kitten. Joe is going to plant the pansy during her lunch hour.

Here is a pansy. Joe is going to swim the PANSY during her lunch hour.

Here is a kitten. Joe is going to swim the PANSY during her lunch hour.

Here is a pansy. Joe is going to swim the pansy during her lunch hour.

Here is a kitten. Joe is going to swim the pansy during her lunch hour.

12. *rattle-carpet*

Here is a rattle. Eason is going to wave the RATTLE at midnight.

Here is a carpet. Eason is going to wave the RATTLE at midnight.

Here is a rattle. Eason is going to wave the rattle at midnight.

Here is a carpet. Eason is going to wave the rattle at midnight.

Here is a rattle. Eason is going to inject the RATTLE at midnight.

Here is a carpet. Eason is going to inject the RATTLE at midnight.

Here is a rattle. Eason is going to inject the rattle at midnight.

Here is a carpet. Eason is going to inject the rattle at midnight.

13. *lifeguard-lifeboat*

Here is a lifeguard. Evelyn is going to swim with the LIFEGUARD in her spare time.

Here is a lifeboat. Evelyn is going to swim with the LIFEGUARD in her spare time.

Here is a lifeguard. Evelyn is going to swim with the lifeguard in her spare time.

Here is a lifeboat. Evelyn is going to swim with the lifeguard in her spare time.

Here is a lifeguard. Evelyn is going to wind the LIFEGUARD in her spare time.

Here is a lifeboat. Evelyn is going to wind the LIFEGUARD in her spare time.

Here is a lifeguard. Evelyn is going to wind the lifeguard in her spare time.

Here is a lifeboat. Evelyn is going to wind the lifeguard in her spare time.

14. *pitcher-turnip*

Here is a pitcher. Jeremy is going to design the PITCHER in an art class.

Here is a turnip. Jeremy is going to design the PITCHER in an art class.

Here is a pitcher. Jeremy is going to design the pitcher in an art class.

Here is a turnip. Jeremy is going to design the pitcher in an art class.

Here is a pitcher. Jeremy is going to write the PITCHER in an art class.

Here is a turnip. Jeremy is going to write the PITCHER in an art class.

Here is a pitcher. Jeremy is going to write the pitcher in an art class.

Here is a turnip. Jeremy is going to write the pitcher in an art class.

15. *hammer-keyboard*

Here is a hammer. Susan is going to pound the HAMMER before the night.

Here is a keyboard. Susan is going to pound the HAMMER before the night.

Here is a hammer. Susan is going to pound the hammer before the night.

Here is a keyboard. Susan is going to pound the hammer before the night.

Here is a hammer. Susan is going to fight the HAMMER before the night.

Here is a keyboard. Susan is going to fight the HAMMER before the night.

Here is a hammer. Susan is going to fight the hammer before the night.

Here is a keyboard. Susan is going to fight the hammer before the night.

16. *lighthouse-lightbulb*

Here is a lighthouse. David is going to operate the LIGHTHOUSE this year.

Here is a lightbulb. David is going to operate the LIGHTHOUSE this year.

Here is a lighthouse. David is going to operate the lighthouse this year.

Here is a lightbulb. David is going to operate the lighthouse this year.

Here is a lighthouse. David is going to admire the LIGHTHOUSE this year.

Here is a lightbulb. David is going to admire the LIGHTHOUSE this year.

Here is a lighthouse. David is going to admire the lighthouse this year.

Here is a lightbulb. David is going to admire the lighthouse this year.

17. *pillar-pillow*

Here is a pillar. Kevin is going to build the PILLAR in one week.

Here is a pillow. Kevin is going to build the PILLAR in one week.

Here is a pillar. Kevin is going to build the pillar in one week.

Here is a pillow. Kevin is going to build the pillar in one week.

Here is a pillar. Kevin is going to run the PILLAR in one week.

Here is a pillow. Kevin is going to run the PILLAR in one week.

Here is a pillar. Kevin is going to run the pillar in one week.

Here is a pillow. Kevin is going to run the pillar in one week.

18. *liquor-lipstick*

Here is a liquor. Stephen is going to taste the LIQUOR at party.

Here is a lipstick. Stephen is going to taste the LIQUOR at party.

Here is a liquor. Stephen is going to taste the liquor at party.

Here is a lipstick. Stephen is going to taste the liquor at party.

Here is a liquor. Stephen is going to interrupt the LIQUOR at party.

Here is a lipstick. Stephen is going to interrupt the LIQUOR at party.

Here is a liquor. Stephen is going to interrupt the liquor at party.

Here is a lipstick. Stephen is going to interrupt the liquor at party.

19. *beetle-collar*

Here is a beetle. Alexander is going to catch the BEETLE after school.

Here is a collar. Alexander is going to catch the BEETLE after school.

Here is a beetle. Alexander is going to catch the beetle after school.

Here is a collar. Alexander is going to catch the beetle after school.

Here is a beetle. Alexander is going to promise the BEETLE after school.

Here is a collar. Alexander is going to promise the BEETLE after school.

Here is a beetle. Alexander is going to promise the beetle after school.

Here is a collar. Alexander is going to promise the beetle after school.

20. *saddle-tuba*

Here is a saddle. Lawrence is going to clean the SADDLE today.

Here is a tuba. Lawrence is going to clean the SADDLE today.

Here is a saddle. Lawrence is going to clean the saddle today.

Here is a tuba. Lawrence is going to clean the saddle today.

Here is a saddle. Lawrence is going to guide the SADDLE today.

Here is a tuba. Lawrence is going to guide the SADDLE today.

Here is a saddle. Lawrence is going to guide the saddle today.

Here is a tuba. Lawrence is going to guide the saddle today.

21. *liver-litter*

Here is a liver. Vincent is going to dissect the LIVER in the biology class.

Here is a litter. Vincent is going to dissect the LIVER in the biology class.

Here is a liver. Vincent is going to dissect the liver in the biology class.

Here is a litter. Vincent is going to dissect the liver in the biology class.

Here is a liver. Vincent is going to catch the LIVER in the biology class.

Here is a litter. Vincent is going to catch the LIVER in the biology class.

Here is a liver. Vincent is going to catch the liver in the biology class.

Here is a litter. Vincent is going to catch the liver in the biology class.

22. *wallet-pigeon*

Here is a wallet. Louis is going to check the WALLET while listening to music.

Here is a pigeon. Louis is going to check the WALLET while listening to music.

Here is a wallet. Louis is going to check the wallet while listening to music.

Here is a pigeon. Louis is going to check the wallet while listening to music.

Here is a wallet. Louis is going to shave the WALLET while listening to music.

Here is a pigeon. Louis is going to shave the WALLET while listening to music.

Here is a wallet. Louis is going to shave the wallet while listening to music.

Here is a pigeon. Louis is going to shave the wallet while listening to music.

23. *candy-walnut*

Here is a candy. Sally is going to pick the CANDY over cocktails.

Here is a walnut. Sally is going to pick the CANDY over cocktails.

Here is a candy. Sally is going to pick the candy over cocktails.

Here is a walnut. Sally is going to pick the candy over cocktails.

Here is a candy. Sally is going to buckle the CANDY over cocktails.

Here is a walnut. Sally is going to buckle the CANDY over cocktails.

Here is a candy. Sally is going to buckle the candy over cocktails.

Here is a walnut. Sally is going to buckle the candy over cocktails.

24. *pendent-pencil*

Here is a pendent. Alice is going to wear the PENDENT before party.

Here is a pencil. Alice is going to wear the PENDENT before party.

Here is a pendent. Alice is going to wear the pendent before party.

Here is a pencil. Alice is going to wear the pendent before party.

Here is a pendent. Alice is going to persuade the PENDENT before party.

Here is a pencil. Alice is going to persuade the PENDENT before party.

Here is a pendent. Alice is going to persuade the pendent before party.

Here is a pencil. Alice is going to persuade the pendent before party.

25. *hammock-whisker*

Here is a hammock. Andrew is going to hang the HAMMOCK before camping.

Here is a whisker. Andrew is going to hang the HAMMOCK before camping.

Here is a hammock. Andrew is going to hang the hammock before camping.

Here is a whisker. Andrew is going to hang the hammock before camping.

Here is a hammock. Andrew is going to irritate the HAMMOCK before camping.

Here is a whisker. Andrew is going to irritate the HAMMOCK before camping.

Here is a hammock. Andrew is going to irritate the hammock before camping.

Here is a whisker. Andrew is going to irritate the hammock before camping.

26. *bandage-banjo*

Here is a bandage. Charlie is going to wrap the BANDAGE before the bedtime.

Here is a banjo. Charlie is going to wrap the BANDAGE before the bedtime.

Here is a bandage. Charlie is going to wrap the bandage before the bedtime.

Here is a banjo. Charlie is going to wrap the bandage before the bedtime.

Here is a bandage. Charlie is going to smear the BANDAGE before the bedtime.

Here is a banjo. Charlie is going to smear the BANDAGE before the bedtime.

Here is a bandage. Charlie is going to smear the bandage before the bedtime.

Here is a banjo. Charlie is going to smear the bandage before the bedtime.

27. *shelving-butter*

Here is a shelving. Michael is going to install the SHELVING on Friday.

Here is a butter. Michael is going to install the SHELVING on Friday.

Here is a shelving. Michael is going to install the shelving on Friday.

Here is a butter. Michael is going to install the shelving on Friday.

Here is a shelving. Michael is going to avoid the SHELVING on Friday.

Here is a butter. Michael is going to avoid the SHELVING on Friday.

Here is a shelving. Michael is going to avoid the shelving on Friday.

Here is a butter. Michael is going to avoid the shelving on Friday.

28. *ladle-sapling*

Here is a ladle. Emily is going to wash the LADLE on Monday of the weeks.

Here is a sapling. Emily is going to wash the LADLE on Monday of the weeks.

Here is a ladle. Emily is going to wash the ladle on Monday of the weeks.

Here is a sapling. Emily is going to wash the ladle on Monday of the weeks.

Here is a ladle. Emily is going to answer the LADLE on Monday of the weeks.

Here is a sapling. Emily is going to answer the LADLE on Monday of the weeks.

Here is a ladle. Emily is going to answer the ladle on Monday of the weeks.

Here is a sapling. Emily is going to answer the ladle on Monday of the weeks.

29. *baker-badger*

Here is a baker. Jessica is going to talk to the BAKER while waiting.

Here is a badger. Jessica is going to talk to the BAKER while waiting.

Here is a baker. Jessica is going to talk to the baker while waiting.

Here is a badger. Jessica is going to talk to the baker while waiting.

Here is a baker. Jessica is going to turn off the BAKER while waiting.

Here is a badger. Jessica is going to turn off the BAKER while waiting.

Here is a baker. Jessica is going to turn off the baker while waiting.

Here is a badger. Jessica is going to turn off the baker while waiting.

30. *cartoon-beaker*

Here is a cartoon. Michelle is going to watch the CARTOON while doing homework.

Here is a beaker. Michelle is going to watch the CARTOON while doing the homework.

Here is a cartoon. Michelle is going to watch the cartoon while doing the homework.

Here is a beaker. Michelle is going to watch the cartoon while doing the homework.

Here is a cartoon. Michelle is going to tickle the CARTOON while doing the homework.

Here is a beaker. Michelle is going to tickle the CARTOON while doing the homework.

Here is a cartoon. Michelle is going to tickle the cartoon while doing the homework.

Here is a beaker. Michelle is going to tickle the cartoon while doing the homework.

31. *kitchen-ketchup*

Here is a kitchen. Mary is going to inspect the KITCHEN today.

Here is a ketchup. Mary is going to inspect the KITCHEN today.

Here is a kitchen. Mary is going to inspect the kitchen today.

Here is a kitchen. Mary is going to inspect the kitchen today.

Here is a ketchup. Mary is going to scold the KITCHEN today.

Here is a kitchen. Mary is going to scold the KITCHEN today.

Here is a ketchup. Mary is going to scold the kitchen today.

Here is a kitchen. Mary is going to scold the kitchen today.

32. *platter-hamper*

Here is a platter. Emma is going to decorate the PLATTER every day.

Here is a hamper. Emma is going to decorate the PLATTER every day.

Here is a platter. Emma is going to decorate the platter every day.

Here is a hamper. Emma is going to decorate the platter every day.

Here is a platter. Emma is going to teach the PLATTER every day.

Here is a hamper. Emma is going to teach the PLATTER every day.

Here is a platter. Emma is going to teach the platter every day.

Here is a hamper. Emma is going to teach the platter every day.

33. *lizard-lily*

Here is a lizard. Jasper is going to feed the LIZARD before work.

Here is a lily. Jasper is going to feed the LIZARD before work.

Here is a lizard. Jasper is going to feed the lizard before work.

Here is a lily. Jasper is going to feed the lizard before work.

Here is a lizard. Jasper is going to cancel the LIZARD before work.

Here is a lily. Jasper is going to cancel the LIZARD before work.

Here is a lizard. Jasper is going to cancel the lizard before work.

Here is a lily. Jasper is going to cancel the lizard before work.

34. *barber-radish*

Here is a barber. Margret is going to have a haircut in the BARBER during the vacation.

Here is a radish. Margret is going to have a haircut in the BARBER during the vacation.

Here is a barber. Margret is going to have a haircut in the barber during the vacation.

Here is a radish. Margret is going to have a haircut in the barber during the vacation.

Here is a barber. Margret is going to unlock the BARBER during the vacation.

Here is a radish. Margret is going to unlock the BARBER during the vacation.

Here is a barber. Margret is going to unlock the barber during the vacation.

Here is a radish. Margret is going to unlock the barber during the vacation.

35. *tulip-corset*

Here is a tulip. Cinderella is going to water the TULIP this morning.

Here is a corset. Cinderella is going to water the TULIP this morning.

Here is a tulip. Cinderella is going to water the tulip this morning.

Here is a corset. Cinderella is going to water the tulip this morning.

Here is a tulip. Cinderella is going to screw in the TULIP this morning.

Here is a corset. Cinderella is going to screw in the TULIP this morning.

Here is a tulip. Cinderella is going to screw in the tulip this morning.

Here is a corset. Cinderella is going to screw in the tulip this morning.

36. *trowel-sandal*

Here is a trowel. Romeo is going to dig with the TROWEL in the garden.

Here is a sandal. Romeo is going to dig with the TROWEL in the garden.

Here is a trowel. Romeo is going to dig with the trowel in the garden.

Here is a sandal. Romeo is going to dig with the trowel in the garden.

Here is a trowel. Romeo is going to promise the TROWEL in the garden.

Here is a sandal. Romeo is going to promise the TROWEL in the garden.

Here is a trowel. Romeo is going to promise the trowel in the garden.

Here is a sandal. Romeo is going to promise the trowel in the garden.

37. *crowbar-panda*

Here is a crowbar. Ernest is going to brandish the CROWBAR before the competition.

Here is a panda. Ernest is going to brandish the CROWBAR before the competition.

Here is a crowbar. Ernest is going to brandish the crowbar before the competition.

Here is a panda. Ernest is going to brandish the crowbar before the competition.

Here is a crowbar. Ernest is going to climb the CROWBAR before the competition.

Here is a panda. Ernest is going to climb the CROWBAR before the competition.

Here is a crowbar. Ernest is going to climb the crowbar before the competition.

Here is a panda. Ernest is going to climb the crowbar before the competition.

38. *lilac-lion*

Here is a lilac. Richard is going to water the LILAC after coming back home.

Here is a lion. Richard is going to water the LILAC after coming back home.

Here is a lilac. Richard is going to water the lilac after coming back home.

Here is a lion. Richard is going to water the lilac after coming back home.

Here is a lilac. Richard is going to write the LILAC after coming back home.

Here is a lion. Richard is going to write the LILAC after coming back home.

Here is a lilac. Richard is going to write the lilac after coming back home.

Here is a lion. Richard is going to write the lilac after coming back home.

39. *jungle-jumper*

Here is a jungle. Martin is going to protect the JUNGLE in the painting class.

Here is a jumper. Martin is going to protect the JUNGLE in the painting class.

Here is a jungle. Martin is going to protect the jungle in the painting class.

Here is a jumper. Martin is going to protect the jungle in the painting class.

Here is a jungle. Martin is going to ask the JUNGLE in the painting class.

Here is a jumper. Martin is going to ask the JUNGLE in the painting class.

Here is a jungle. Martin is going to ask the jungle in the painting class.

Here is a jumper. Martin is going to ask the jungle in the painting class.

40. *turkey-hamster*

Here is a turkey. Brian is going to fill the TURKEY before the Thanksgiving day.

Here is a hamster. Brian is going to fill the TURKEY before the Thanksgiving day.

Here is a turkey. Brian is going to fill the turkey before the Thanksgiving day.

Here is a hamster. Brian is going to fill the turkey before the Thanksgiving day.

Here is a turkey. Brian is going to switch the TURKEY before the Thanksgiving day.

Here is a hamster. Brian is going to switch the TURKEY before the Thanksgiving day.

Here is a turkey. Brian is going to switch the turkey before the Thanksgiving day.

Here is a hamster. Brian is going to switch the turkey before the Thanksgiving day.

C.2 Mandarin stimuli

For Mandarin stimuli, targets accentuated with contrastive pitch accent are in "【】."

C.2.1 Critical sentences

1.苹果-瓶子*"apple-bottle"*

你可以看到一些苹果。现在丽丽要去削了【苹果】来给自己的好朋友吃。

你可以看到一些瓶子。现在丽丽要去削了【苹果】来给自己的好朋友吃。

你可以看到一些苹果。现在丽丽要去削了苹果来给自己的好朋友吃。

你可以看到一些瓶子。现在丽丽要去削了苹果来给自己的好朋友吃。

你可以看到一些苹果。现在丽丽要去拥抱【苹果】来给自己的好朋友吃。

你可以看到一些瓶子。现在丽丽要去拥抱【苹果】来给自己的好朋友吃。

你可以看到一些苹果。现在丽丽要去拥抱苹果来给自己的好朋友吃。

你可以看到一些瓶子。现在丽丽要去拥抱苹果来给自己的好朋友吃。

2. 橙汁-城堡*"orange juice-castle"*

你可以看到一些橙汁。现在安迪要去榨了【橙汁】来准备明天的早餐。

你可以看到一些城堡。现在安迪要去榨了【橙汁】来准备明天的早餐。

你可以看到一些橙汁。现在安迪要去榨了橙汁来准备明天的早餐。

你可以看到一些城堡。现在安迪要去榨了橙汁来准备明天的早餐。

你可以看到一些橙汁。现在安迪要去忘记【橙汁】来准备明天的早餐。

你可以看到一些城堡。现在安迪要去忘记【橙汁】来准备明天的早餐。

你可以看到一些橙汁。现在安迪要去忘记橙汁来准备明天的早餐。

你可以看到一些城堡。现在安迪要去忘记橙汁来准备明天的早餐。

3. 蜡烛-辣椒*"candle-pepper"*

你可以看到一个蜡烛。现在吉娜要去点燃【蜡烛】才能看书。

你可以看到一些辣椒。现在吉娜要去点燃【蜡烛】才能看书。

你可以看到一个蜡烛。现在吉娜要去点燃蜡烛才能看书。

你可以看到一些辣椒。现在吉娜要去点燃蜡烛才能看书。

你可以看到一个蜡烛。现在吉娜要去吃了【蜡烛】才能看书。

你可以看到一些辣椒。现在吉娜要去吃了【蜡烛】才能看书。

你可以看到一个蜡烛。现在吉娜要去吃了蜡烛才能看书。

你可以看到一些辣椒。现在吉娜要去吃了蜡烛才能看书。

4.大米-大树*"rice-tree"*

你可以看到一些大米。现在小陈要去煮熟【大米】来做晚饭。

你可以看到一些大树。现在小陈要去煮熟【大米】来做晚饭。

你可以看到一些大米。现在小陈要去煮熟大米来做晚饭。

你可以看到一些大树。现在小陈要去煮熟大米来做晚饭。

你可以看到一些大米。现在小陈要去租了【大米】来做晚饭。

你可以看到一些大树。现在小陈要去租了【大米】来做晚饭。

你可以看到一些大米。现在小陈要去租了大米来做晚饭。

你可以看到一些大树。现在小陈要去租了大米来做晚饭。

5.大米-大树*"rice-tree"*

你可以看到一个香水。现在姗姗要去喷了【香水】再去约会。

你可以看到一些香蕉。现在姗姗要去喷了【香水】再去约会。

你可以看到一个香水。现在姗姗要去喷了香水再去约会。

你可以看到一些香蕉。现在姗姗要去喷了香水再去约会。

你可以看到一个香水。现在姗姗要去打印【香水】再去约会。

你可以看到一些香蕉。现在姗姗要去打印【香水】再去约会。

你可以看到一个香水。现在姗姗要去打印香水再去约会。

你可以看到一些香蕉。现在姗姗要去打印香水再去约会。

6.钢琴-钢铁*"piano-steel"*

你可以看到一个钢琴。现在爱丽丝要去弹奏【钢琴】给老师检查练习。

你可以看到一些钢铁。现在爱丽丝要去弹奏【钢琴】给老师检查练习。

你可以看到一个钢琴。现在爱丽丝要去弹奏钢琴给老师检查练习。

你可以看到一些钢铁。现在爱丽丝要去弹奏钢琴给老师检查练习。

你可以看到一个钢琴。现在爱丽丝要去帮助【钢琴】给老师检查练习。

你可以看到一些钢铁。现在爱丽丝要去帮助【钢琴】给老师检查练习。

你可以看到一个钢琴。现在爱丽丝要去帮助钢琴给老师检查练习。

你可以看到一些钢铁。现在爱丽丝要去帮助钢琴给老师检查练习。

7.飞机-飞鸟*"plane-flying bird"*

你可以看到一个飞机。现在丹妮要去乘坐【飞机】去日本度假。

你可以看到一些飞鸟。现在丹妮要去乘坐【飞机】去日本度假。

你可以看到一个飞机。现在丹妮要去乘坐飞机去日本度假。

你可以看到一些飞鸟。现在丹妮要去乘坐飞机去日本度假。

你可以看到一个飞机。现在丹妮要去污染【飞机】去日本度假。

你可以看到一些飞鸟。现在丹妮要去污染【飞机】去日本度假。

你可以看到一个飞机。现在丹妮要去污染飞机去日本度假。

你可以看到一些飞鸟。现在丹妮要去污染飞机去日本度假。

8.火车-火鸡*"train-turkey"*

你可以看到一个火车。现在小周要去乘坐【火车】去上班。

你可以看到一个火鸡。现在小周要去乘坐【火车】去上班。

你可以看到一个火车。现在小周要去乘坐火车去上班。

你可以看到一个火鸡。现在小周要去乘坐火车去上班。

你可以看到一个火车。现在小周要去打开【火车】去上班。

你可以看到一个火鸡。现在小周要去打开【火车】去上班。

你可以看到一个火车。现在小周要去打开火车去上班。

你可以看到一个火鸡。现在小周要去打开火车去上班。

9.毛衣-毛笔*"sweater-writing brush"*

你可以看到一些毛衣。现在薇薇要去织了【毛衣】再去睡觉。

你可以看到一些毛笔。现在薇薇要去织了【毛衣】再去睡觉。

你可以看到一些毛衣。现在微微要去织了毛衣再去睡觉。

你可以看到一些毛笔。现在微微要去织了毛衣再去睡觉。

你可以看到一些毛衣。现在薇薇要去减少【毛衣】再去睡觉。

你可以看到一些毛笔。现在薇薇要去减少【毛衣】再去睡觉。

你可以看到一些毛衣。现在微微要去减少毛衣再去睡觉。

你可以看到一些毛笔。现在微微要去减少毛衣再去睡觉。

10.水稻-水獭``rice crop-otter"

你可以看到一些水稻。现在小王要去种植【水稻】来赚钱。

你可以看到一些水獭。现在小王要去种植【水稻】来赚钱。

你可以看到一些水稻。现在小王要去种植水稻来赚钱。

你可以看到一些水獭。现在小王要去种植水稻来赚钱。

你可以看到一些水稻。现在小王要去陪伴【水稻】来赚钱。

你可以看到一些水獭。现在小王要去陪伴【水稻】来赚钱。

你可以看到一些水稻。现在小王要去陪伴水稻来赚钱。

你可以看到一些水獭。现在小王要去陪伴水稻来赚钱。

11.玫瑰-煤炭*"rose-coal"*

你可以看到一些玫瑰。现在慧文要去种植【玫瑰】来美化花园。

你可以看到一些煤炭。现在慧文要去种植【玫瑰】来美化花园。

你可以看到一些玫瑰。现在慧文要去种植玫瑰来美化花园。

你可以看到一些煤炭。现在慧文要去种植玫瑰来美化花园。

你可以看到一些玫瑰。现在慧文要去联系【玫瑰】来美化花园。

你可以看到一些煤炭。现在慧文要去联系【玫瑰】来美化花园。

你可以看到一些玫瑰。现在慧文要去联系玫瑰来美化花园。

你可以看到一些煤炭。现在慧文要去联系玫瑰来美化花园。

12.啤酒-皮鞋*"beer-leather shoes"*

你可以看到一些啤酒。现在书贤要去喝了【啤酒】再去吃饭。

你可以看到一个皮鞋。现在书贤要去喝了【啤酒】再去吃饭。

你可以看到一些啤酒。现在书贤要去喝了啤酒再去吃饭。

你可以看到一个皮鞋。现在书贤要去喝了啤酒再去吃饭。

你可以看到一些啤酒。现在书贤要去玩了【啤酒】再去吃饭。

你可以看到一个皮鞋。现在书贤要去玩了【啤酒】再去吃饭。

你可以看到一些啤酒。现在书贤要去玩了啤酒再去吃饭。

你可以看到一个皮鞋。现在书贤要去玩了啤酒再去吃饭。

13.老师-老鼠*"teacher-rat"*

你可以看到一个老师。现在晓彤要去拜访【老师】因为很多年没有见了。

你可以看到一些老鼠。现在晓彤要去拜访【老师】因为很多年没有见了。

你可以看到一个老师。现在晓彤要去拜访老师因为很多年没有见了。

你可以看到一些老鼠。现在晓彤要去拜访老师因为很多年没有见了。

你可以看到一个老师。现在晓彤要去推迟【老师】因为很多年没有见了。

你可以看到一些老鼠。现在晓彤要去推迟【老师】因为很多年没有见了。

你可以看到一个老师。现在晓彤要去推迟老师因为很多年没有见了。

你可以看到一些老鼠。现在晓彤要去推迟老师因为很多年没有见了。

14.口红-口琴*"lipstick-harmonica"*

你可以看到一些口红。现在晓霞要去涂了【口红】再去开会。

你可以看到一个口琴。现在晓霞要去涂了【口红】再去开会。

你可以看到一些口红。现在晓霞要去涂了口红再去开会。

你可以看到一个口琴。现在晓霞要去涂了口红再去开会。

你可以看到一些口红。现在晓霞要去签了【口红】再去开会。

你可以看到一个口琴。现在晓霞要去签了【口红】再去开会。

你可以看到一些口红。现在晓霞要去签了口红再去开会。

你可以看到一个口琴。现在晓霞要去签了口红再去开会。

15.词典-瓷器*"dictionary-porcelain"*

你可以看到一个词典。现在小涵要去查阅【词典】才能写好这篇文章。

你可以看到一个瓷器。现在小涵要去查阅【词典】才能写好这篇文章。

你可以看到一个词典。现在小涵要去查阅词典才能写好这篇文章。

你可以看到一个瓷器。现在小涵要去查阅词典才能写好这篇文章。

你可以看到一个词典。现在小涵要去玩了【词典】才能写好这篇文章。

你可以看到一个瓷器。现在小涵要去玩了【词典】才能写好这篇文章。

你可以看到一个词典。现在小涵要去玩了词典才能写好这篇文章。

你可以看到一个瓷器。现在小涵要去玩了词典才能写好这篇文章。

16.课本-课桌*"textbook-desk"*

你可以看到一个课本. 现在丽丽要去复习【课本】才能在这次考试中考得好。

你可以看到一个课桌. 现在丽丽要去复习【课本】才能在这次考试中考得好。

你可以看到一个课本. 现在丽丽要去复习课本才能在这次考试中考得好。

你可以看到一个课桌. 现在丽丽要去复习课本才能在这次考试中考得好。

你可以看到一个课本. 现在丽丽要去违反【课本】才能在这次考试中考得好。

你可以看到一个课桌. 现在丽丽要去违反【课本】才能在这次考试中考得好。

你可以看到一个课本. 现在丽丽要去违反课本才能在这次考试中考得好。

你可以看到一个课桌. 现在丽丽要去违反课本才能在这次考试中考得好。

17. 地铁-地毯*"metro-carpet"*

你可以看到一个地铁。现在小欢要去乘坐【地铁】再去逛街。

你可以看到一个地毯。现在小欢要去乘坐【地铁】再去逛街。

你可以看到一个地铁。现在丽丽要去乘坐地铁再去逛街。

你可以看到一个地毯。现在丽丽要去乘坐地铁再去逛街。

你可以看到一个地铁。现在小欢要去送了【地铁】再去逛街。

你可以看到一个地毯。现在小欢要去送了【地铁】再去逛街。

你可以看到一个地铁。现在丽丽要去送了地铁再去逛街。

你可以看到一个地毯。现在丽丽要去送了地铁再去逛街。

18. 食物-石头*"food-stone"*

你可以看到一个食物。现在丽丽要去储存【食物】因为现在不容易出门。

你可以看到一个石头。现在丽丽要去储存【食物】因为现在不容易出门。

你可以看到一个食物。现在丽丽要去储存食物因为现在不容易出门。

你可以看到一个石头。现在丽丽要去储存食物因为现在不容易出门。

你可以看到一个食物。现在丽丽要去忘记【食物】因为现在不容易出门。

你可以看到一个石头。现在丽丽要去忘记【食物】因为现在不容易出门。

你可以看到一个食物。现在丽丽要去忘记食物因为现在不容易出门。

你可以看到一个石头。现在丽丽要去忘记食物因为现在不容易出门。

19. 医院-衣服*"hospital-cloth"*

你可以看到一个医院。现在丽丽要去前往【医院】看病。

你可以看到一个衣服。现在丽丽要去前往【医院】看病。

你可以看到一个医院。现在丽丽要去前往医院看病。

你可以看到一个衣服。现在丽丽要去前往医院看病。

你可以看到一个医院。现在丽丽要去栽了【医院】看病。

你可以看到一个衣服。现在丽丽要去栽了【医院】看病。

你可以看到一个医院。现在丽丽要去栽了医院看病。

你可以看到一个衣服。现在丽丽要去栽了医院看病。

20. 空调-空姐*"air conditioner-air stewardess"*

你可以看到一个空调。现在丽丽要去安装【空调】这样夏天就变得凉快了。

你可以看到一个空姐。现在丽丽要去安装【空调】这样夏天就变得凉快了。

你可以看到一个空调。现在丽丽要去安装空调这样夏天就变得凉快了。

你可以看到一个空姐。现在丽丽要去安装空调这样夏天就变得凉快了。

你可以看到一个空调。现在丽丽要去问了【空调】这样夏天就变得凉快了。

你可以看到一个空姐。现在丽丽要去问了【空调】这样夏天就变得凉快了。

你可以看到一个空调。现在丽丽要去问了空调这样夏天就变得凉快了。

你可以看到一个空姐。现在丽丽要去问了空调这样夏天就变得凉快了。

21. 篮球-篮子*"basketball-basket"*

你可以看到一个篮球。现在丽丽要去打了【篮球】来减肥。

你可以看到一个篮子。现在丽丽要去打了【篮球】来减肥。

你可以看到一个篮球。现在丽丽要去打了篮球来减肥。

你可以看到一个篮子。现在丽丽要去打了篮球来减肥。

你可以看到一个篮球。现在丽丽要去负责【篮球】来减肥。

你可以看到一个篮子。现在丽丽要去负责【篮球】来减肥。

你可以看到一个篮球。现在丽丽要去负责篮球来减肥。

你可以看到一个篮子。现在丽丽要去负责篮球来减肥。

22. 电视-电器*"Television-Electrical appliances"*

你可以看到一个电视。现在丽丽要去打开【电视】度过美好的夜晚。

你可以看到一个电器。现在丽丽要去打开【电视】度过美好的夜晚。

你可以看到一个电视。现在丽丽要去打开电视度过美好的夜晚。

你可以看到一个电器。现在丽丽要去打开电视度过美好的夜晚。

你可以看到一个电视。现在丽丽要去下载【电视】度过美好的夜晚。

你可以看到一个电器。现在丽丽要去下载【电视】度过美好的夜晚。

你可以看到一个电视。现在丽丽要去下载电视度过美好的夜晚。

你可以看到一个电器。现在丽丽要去下载电视度过美好的夜晚。

23. 书橱-书法*"shelving-calligraphy"*

你可以看到一个书橱。现在丽丽要去练习【书法】才能在比赛中拿到好成绩。

你可以看到一个书法。现在丽丽要去练习【书法】才能在比赛中拿到好成绩。

你可以看到一个书橱。现在丽丽要去练习书法才能在比赛中拿到好成绩。

你可以看到一个书法。现在丽丽要去搬走书法才能在比赛中拿到好成绩。

你可以看到一个书橱。现在丽丽要去搬走【书法】才能在比赛中拿到好成绩。

你可以看到一个书法。现在丽丽要去搬走【书法】才能在比赛中拿到好成绩。

你可以看到一个书橱。现在丽丽要去搬走书法才能在比赛中拿到好成绩。

24.树苗-树懒*"sapling-sloth"*

你可以看到一个树苗。现在丽丽要去培育【树苗】来美化环境。

你可以看到一个树懒。现在丽丽要去培育【树苗】来美化环境。

你可以看到一个树苗。现在丽丽要去培育树苗来美化环境。

你可以看到一个树懒。现在丽丽要去培育树苗来美化环境。

你可以看到一个树苗。现在丽丽要去维修【树苗】来美化环境。

你可以看到一个树懒。现在丽丽要去维修【树苗】来美化环境。

你可以看到一个树苗。现在丽丽要去维修树苗来美化环境。

你可以看到一个树懒。现在丽丽要去维修树苗来美化环境。

25. 灯泡-灯塔*"light bulb-lighthouse"*

你可以看到一个灯泡。现在小香要去更换【灯泡】因为老的坏掉了。

你可以看到一个灯塔。现在小香要去更换【灯泡】因为老的坏掉了。

你可以看到一个灯泡。现在小香要去更换灯泡因为老的坏掉了。

你可以看到一个灯塔。现在小香要去更换灯泡因为老的坏掉了。

你可以看到一个灯泡。现在小香要去邀请【灯泡】因为老的坏掉了。

你可以看到一个灯塔。现在小香要去邀请【灯泡】因为老的坏掉了。

你可以看到一个灯泡。现在小香要去邀请灯泡因为老的坏掉了。

你可以看到一个灯塔。现在小香要去邀请灯泡因为老的坏掉了。

26. 铅笔-铅球*"pencil-shot put"*

你可以看到一个铅笔。现在小萱要去削了【铅笔】再去上美术课。

你可以看到一个铅球。现在小萱要去削了【铅笔】再去上美术课。

你可以看到一个铅笔。现在小萱要去削了铅笔再去上美术课。

你可以看到一个铅球。现在小萱要去削了铅笔再去上美术课。

你可以看到一个铅笔。现在小萱要去骂了【铅笔】再去上美术课。

你可以看到一个铅球。现在小萱要去骂了【铅笔】再去上美术课。

你可以看到一个铅笔。现在小萱要去骂了铅笔再去上美术课。

你可以看到一个铅球。现在小萱要去骂了铅笔再去上美术课。

27. 凉鞋-凉席*"sandal-mat"*

你可以看到一个凉鞋。现在小思要去穿上【凉鞋】去逛街。

你可以看到一个凉席。现在小思要去穿上【凉鞋】去逛街。

你可以看到一个凉鞋。现在小思要去穿上凉鞋去逛街。

你可以看到一个凉席。现在小思要去穿上凉鞋去逛街。

你可以看到一个凉鞋。现在小思要去练习【凉鞋】去逛街。

你可以看到一个凉席。现在小思要去练习【凉鞋】去逛街。

你可以看到一个凉鞋。现在小思要去练习凉鞋去逛街。

你可以看到一个凉席。现在小思要去练习凉鞋去逛街。

28. 卡车-卡通*"truck-cartoon"*

你可以看到一个卡车。现在李大爷要去乘坐【卡车】回家。

你可以看到一个卡通。现在李大爷要去乘坐【卡车】回家。

你可以看到一个卡车。现在李大爷要去乘坐卡车回家。

你可以看到一个卡通。现在李大爷要去乘坐卡车回家。

你可以看到一个卡车。现在李大爷要去喝了【卡车】回家。

你可以看到一个卡通。现在李大爷要去喝了【卡车】回家。

你可以看到一个卡车。现在李大爷要去喝了卡车回家。

你可以看到一个卡通。现在李大爷要去喝了卡车回家。

29. 花瓶-花生*"vase-peanut"*

你可以看到一个花瓶。现在小蔚要去洗了【花瓶】然后再去睡觉。

你可以看到一个花生。现在小蔚要去洗了【花瓶】然后再去睡觉。

你可以看到一个花瓶。现在小蔚要去洗了花瓶然后再去睡觉。

你可以看到一个花生。现在小蔚要去洗了花瓶然后再去睡觉。

你可以看到一个花瓶。现在小蔚要去爬了【花瓶】然后再去睡觉。

你可以看到一个花生。现在小蔚要去爬了【花瓶】然后再去睡觉。

你可以看到一个花瓶。现在小蔚要去爬了花瓶然后再去睡觉。

你可以看到一个花生。现在小蔚要去爬了花瓶然后再去睡觉。

30. 小吃-小猫*"snacks-kitten"*

你可以看到一个小吃。现在小芙要去吃了【小吃】再回家看电影。

你可以看到一个小猫。现在小芙要去吃了【小吃】再回家看电影。

你可以看到一个小吃。现在小芙要去吃了小吃再回家看电影。

你可以看到一个小猫。现在小芙要去吃了小吃再回家看电影。

你可以看到一个小吃。现在小芙要去交了【小吃】再回家看电影。

你可以看到一个小猫。现在小芙要去交了【小吃】再回家看电影。

你可以看到一个小吃。现在小芙要去交了小吃再回家看电影。

你可以看到一个小猫。现在小芙要去交了小吃再回家看电影。

31. 报纸-鲍鱼*"newspaper-abalone"*

你可以看到一个报纸。现在钱奶奶要去阅读【报纸】来了解新闻。

你可以看到一个鲍鱼。现在钱奶奶要去阅读【报纸】来了解新闻。

你可以看到一个报纸。现在钱奶奶要去阅读报纸来了解新闻。

你可以看到一个鲍鱼。现在钱奶奶要去阅读报纸来了解新闻。

你可以看到一个报纸。现在钱奶奶要去锻炼【报纸】来了解新闻。

你可以看到一个鲍鱼。现在钱奶奶要去锻炼【报纸】来了解新闻。

你可以看到一个报纸。现在钱奶奶要去锻炼报纸来了解新闻。

你可以看到一个鲍鱼。现在钱奶奶要去锻炼报纸来了解新闻。

32. 手机-手袋*"phone-handbag"*

你可以看到一个手机。现在小蕴要玩了【手机】再去学习。

你可以看到一个手袋。现在小蕴要玩了【手机】再去学习。

你可以看到一个手机。现在小蕴要玩了手机再去学习。

你可以看到一个手袋。现在小蕴要玩了手机再去学习。

你可以看到一个手机。现在小蕴要推迟【手机】再去学习。

你可以看到一个手袋。现在小蕴要推迟【手机】再去学习。

你可以看到一个手机。现在小蕴要推迟手机再去学习。

你可以看到一个手袋。现在小蕴要推迟手机再去学习。

33. 面条-面包*"noodle-bread"*

你可以看到一个面条。现在小芹要去煮上【面条】再去看电影。

你可以看到一个面包。现在小芹要去煮上【面条】再去看电影。

你可以看到一个面条。现在小芹要去煮上面条再去看电影。

你可以看到一个面包。现在小芹要去煮上面条再去看电影。

你可以看到一个面条。现在小芹要去穿上【面条】再去看电影。

你可以看到一个面包。现在小芹要去穿上【面条】再去看电影。

你可以看到一个面条。现在小芹要去穿上面条再去看电影。

你可以看到一个面包。现在小芹要去穿上面条再去看电影。

34. 牛奶-牛肉 "*milk-beef*"

你可以看到一个牛奶。现在小琳要去喝了【牛奶】再吃晚饭。

你可以看到一个牛肉。现在小琳要去喝了【牛奶】再吃晚饭。

你可以看到一个牛奶。现在小琳要去喝了牛奶再吃晚饭。

你可以看到一个牛肉。现在小琳要去喝了牛奶再吃晚饭。

你可以看到一个牛奶。现在小琳要去表演【牛奶】再吃晚饭。

你可以看到一个牛肉。现在小琳要去表演【牛奶】再吃晚饭。

你可以看到一个牛奶。现在小琳要去表演牛奶再吃晚饭。

你可以看到一个牛肉。现在小琳要去表演牛奶再吃晚饭。

35. 风扇-蜂蜜*"fan-honey"*

你可以看到一个风扇。现在小瑶要去吹了【风扇】才没有那么热。

你可以看到一个蜂蜜。现在小瑶要去吹了【风扇】才没有那么热。

你可以看到一个风扇。现在小瑶要去吹了风扇才没有那么热。

你可以看到一个蜂蜜。现在小瑶要去吹了风扇才没有那么热。

你可以看到一个风扇。现在小瑶要去参加【风扇】才没有那么热。

你可以看到一个蜂蜜。现在小瑶要去参加【风扇】才没有那么热。

你可以看到一个风扇。现在小瑶要去参加风扇才没有那么热。

你可以看到一个蜂蜜。现在小瑶要去参加风扇才没有那么热。

36. '电线-电脑'*wire-computer"*

你可以看到一个电线。现在小宜要去剪断【电线】因为太长了。

你可以看到一个电脑。现在小宜要去剪断【电线】因为太长了。

你可以看到一个电线。现在小宜要去剪断电线因为太长了。

你可以看到一个电脑。现在小宜要去剪断电线因为太长了。

你可以看到一个电线。现在小宜要去跳了【电线】因为太长了。

你可以看到一个电脑。现在小宜要去跳了【电线】因为太长了。

你可以看到一个电线。现在小宜要去跳了电线因为太长了。

你可以看到一个电脑。现在小宜要去跳了电线因为太长了。

37. 酒精-酒杯*"surgical-spirit"*

你可以看到一个酒精。现在小宜要去用了【酒精】消毒之后再吃饭。

你可以看到一个酒杯。现在小宜要去用了【酒精】消毒之后再吃饭。

你可以看到一个酒精。现在小宜要去用了酒精消毒之后再吃饭。

你可以看到一个酒杯。现在小宜要去用了酒精消毒之后再吃饭。

你可以看到一个酒精。现在小宜要去修了【酒精】消毒之后再吃饭。

你可以看到一个酒杯。现在小宜要去修了【酒精】消毒之后再吃饭。

你可以看到一个酒精。现在小宜要去修了酒精消毒之后再吃饭。

你可以看到一个酒杯。现在小宜要去修了酒精消毒之后再吃饭。

38. 冰箱-冰块*"refrigerator-ice cubes"*

你可以看到一个冰箱。现在小雯要去清洁【冰箱】来保持卫生。

你可以看到一个冰块。现在小雯要去清洁【冰箱】来保持卫生。

你可以看到一个冰箱。现在小雯要去清洁冰箱来保持卫生。

你可以看到一个冰块。现在小雯要去清洁冰箱来保持卫生。

你可以看到一个冰箱。现在小雯要去摔了【冰箱】来保持卫生。

你可以看到一个冰块。现在小雯要去摔了【冰箱】来保持卫生。

你可以看到一个冰箱。现在小雯要去摔了冰箱来保持卫生。

你可以看到一个冰块。现在小雯要去摔了冰箱来保持卫生。

39. 雨伞-雨衣*"umbrella-rain coat"*

你可以看到一个雨伞。现在小冰要去撑着【雨伞】才能去上学。

你可以看到一个雨衣。现在小冰要去撑着【雨伞】才能去上学。

你可以看到一个雨伞。现在小冰要去撑着雨伞才能去上学。

你可以看到一个雨衣。现在小冰要去撑着雨伞才能去上学。

你可以看到一个雨伞。现在小冰要去帮了【雨伞】才能去上学。

你可以看到一个雨衣。现在小冰要去帮了【雨伞】才能去上学。

你可以看到一个雨伞。现在小冰要去帮了雨伞才能去上学。

你可以看到一个雨衣。现在小冰要去帮了雨伞才能去上学。

40. 乌龟-乌梅*"tortoise-black plum"*

你可以看到一个乌龟。现在小枫要去喂养【乌龟】因为别的都不好养。

你可以看到一个乌梅。现在小枫要去喂养【乌龟】因为别的都不好养。

你可以看到一个乌龟。现在小枫要去喂养乌龟因为别的都不好养。

你可以看到一个乌梅。现在小枫要去喂养乌龟因为别的都不好养。

你可以看到一个乌龟。现在小枫要去检查【乌龟】因为别的都不好养。

你可以看到一个乌梅。现在小枫要去检查【乌龟】因为别的都不好养。

你可以看到一个乌龟。现在小枫要去检查乌龟因为别的都不好养。

你可以看到一个乌梅。现在小枫要去检查乌龟因为别的都不好养。

C.2.2 Filler sentences

1. 瓶子-卡车*"bottle-truck"*

你可以看到一个瓶子。现在小彤要去洗了【瓶子】再去上班。

你可以看到一个卡车。现在小彤要去洗了【瓶子】再去上班。

你可以看到一个瓶子。现在小彤要去洗了瓶子再去上班。

你可以看到一个卡车。现在小彤要去洗了瓶子再去上班。

你可以看到一个瓶子。现在小彤要去回答【瓶子】再去上班。

你可以看到一个卡车。现在小彤要去回答【瓶子】再去上班。

你可以看到一个瓶子。现在小彤要去回答瓶子再去上班。

你可以看到一个卡车。现在小彤要去回答瓶子再去上班。

2. 城堡-蜡烛*"castle-candle"*

你可以看到一个城堡。现在小清要去访问【城堡】再准备早餐。

你可以看到一个蜡烛。现在小清要去访问【城堡】再准备早餐。

你可以看到一个城堡。现在小清要去访问城堡再准备早餐。

你可以看到一个蜡烛。现在小清要去访问城堡再准备早餐。

你可以看到一个城堡。现在小清要去切了【城堡】再准备早餐。

你可以看到一个蜡烛。现在小清要去切了【城堡】再准备早餐。

你可以看到一个城堡。现在小清要去切了城堡再准备早餐。

你可以看到一个蜡烛。现在小清要去切了城堡再准备早餐。

3. 辣椒-毛衣*"pepper-sweater"*

你可以看到一个辣椒。现在小旭要去品尝【辣椒】因为这是当地的特色。

你可以看到一个毛衣。现在小旭要去品尝【辣椒】因为这是当地的特色。

你可以看到一个辣椒。现在小旭要去品尝辣椒因为这是当地的特色。

你可以看到一个毛衣。现在小旭要去品尝辣椒因为这是当地的特色。

你可以看到一个辣椒。现在小旭要去遵守【辣椒】因为这是当地的特色。

你可以看到一个毛衣。现在小旭要去遵守【辣椒】因为这是当地的特色。

你可以看到一个辣椒。现在小旭要去遵守辣椒因为这是当地的特色。

你可以看到一个毛衣。现在小旭要去遵守辣椒因为这是当地的特色。

4. 大树-香水*"tree-perfume"*

你可以看到一个大树。现在老赵要去砍【大树】然后拿去卖。

你可以看到一个香水。现在老赵要去砍【大树】然后拿去卖。

你可以看到一个大树。现在老赵要去砍大树然后拿去卖。

你可以看到一个香水。现在老赵要去砍大树然后拿去卖。

你可以看到一个大树。现在老赵要去玩【大树】然后拿去卖。

你可以看到一个香水。现在老赵要去玩【大树】然后拿去卖。

你可以看到一个大树。现在老赵要去玩大树然后拿去卖。

你可以看到一个香水。现在老赵要去玩大树然后拿去卖。

5. 香蕉-钢琴*"banana-piano"*

你可以看到一个香蕉。现在小楠要去吃【香蕉】来减肥。

你可以看到一个钢琴。现在小楠要去吃【香蕉】来减肥。

你可以看到一个香蕉。现在小楠要去吃香蕉来减肥。

你可以看到一个钢琴。现在小楠要去吃香蕉来减肥。

你可以看到一个香蕉。现在小楠要去避免【香蕉】来减肥。

你可以看到一个钢琴。现在小楠要去避免【香蕉】来减肥。

你可以看到一个香蕉。现在小楠要去避免香蕉来减肥。

你可以看到一个钢琴。现在小楠要去避免香蕉来减肥。

6. 钢铁-橙汁*"steel-orange juice"*

你可以看到一些钢铁。现在小秀要去买【钢铁】来建房子。

你可以看到一些橙汁。现在小秀要去买【钢铁】来建房子。

你可以看到一些钢铁。现在小秀要去买钢铁来建房子。

你可以看到一些橙汁。现在小秀要去买钢铁来建房子。

你可以看到一些钢铁。现在小秀要去修改【钢铁】来建房子。

你可以看到一些橙汁。现在小秀要去修改【钢铁】来建房子。

你可以看到一些钢铁。现在小秀要去修改钢铁来建房子。

你可以看到一些橙汁。现在小秀要去修改钢铁来建房子。

7. 大米-飞鸟*"rice-flying bird"*

你可以看到一个大米。现在小嘉要去拍下【飞鸟】留作纪念。

你可以看到一个飞鸟。现在小嘉要去拍下【飞鸟】留作纪念。

你可以看到一个大米。现在小嘉要去拍下飞鸟留作纪念。

你可以看到一个飞鸟。现在小嘉要去离开飞鸟留作纪念。

你可以看到一个大米。现在小嘉要去离开【飞鸟】留作纪念。

你可以看到一个飞鸟。现在小嘉要去离开【飞鸟】留作纪念。

你可以看到一个大米。现在小嘉要去离开飞鸟留作纪念。

你可以看到一个大米。现在小嘉要去离开飞鸟留作纪念。

8. 火鸡-飞机*"turkey-plane"*

你可以看到一个火鸡。现在小妮要去烤【火鸡】来庆祝自己的生日。

你可以看到一个飞机。现在小妮要烤【火鸡】来庆祝自己的生日。

你可以看到一个火鸡。现在小妮要烤火鸡来庆祝自己的生日。

你可以看到一个飞机。现在小妮要烤火鸡来庆祝自己的生日。

你可以看到一个火鸡。现在小妮要限制【火鸡】来庆祝自己的生日。

你可以看到一个飞机。现在小妮要限制【火鸡】来庆祝自己的生日。

你可以看到一个火鸡。现在小妮要限制火鸡来庆祝自己的生日。

你可以看到一个飞机。现在小妮要限制火鸡来庆祝自己的生日。

9. 毛笔-火车*"writing brush-train"*

你可以看到一个毛笔。现在小琼要去用【毛笔】去写字。

你可以看到一个火车。现在小琼要用【毛笔】去写字。

你可以看到一个毛笔。现在小琼要用毛笔去写字。

你可以看到一个火车。现在小琼要用毛笔去写字。

你可以看到一个毛笔。现在小琼要保持【毛笔】去写字。

你可以看到一个火车。现在小琼要保持【毛笔】去写字。

你可以看到一个毛笔。现在小琼要保持毛笔去写字。

你可以看到一个火车。现在小琼要保持毛笔去写字。

10. 水獭-小吃*“otter-snacks”*

你可以看到一个水獭。现在小檀要去喂养【水獭】因为他是动物管理员。

你可以看到一个小吃。现在小檀要去喂养【水獭】因为他是动物管理员。

你可以看到一个水獭。现在小檀要去喂养水獭因为他是动物管理员。

你可以看到一个小吃。现在小檀要去喂养水獭因为他是动物管理员。

你可以看到一个水獭。现在小檀要去拜访【水獭】因为他是动物管理员。

你可以看到一个小吃。现在小檀要去拜访【水獭】因为他是动物管理员。

你可以看到一个水獭。现在小檀要去拜访水獭因为他是动物管理员。

你可以看到一个小吃。现在小檀要去拜访水獭因为他是动物管理员。

11. 煤炭-水稻*“coal-rice crop”*

你可以看到一个煤炭。现在小盈要去卖【煤炭】来赚钱。

你可以看到一个水稻。现在小盈要去卖【煤炭】来赚钱。

你可以看到一个煤炭。现在小盈要去卖煤炭来赚钱。

你可以看到一个水稻。现在小盈要去卖煤炭来赚钱。

你可以看到一个煤炭。现在小盈要去骗【煤炭】来赚钱。

你可以看到一个水稻。现在小盈要去骗【煤炭】来赚钱。

你可以看到一个煤炭。现在小盈要去骗煤炭来赚钱。

你可以看到一个水稻。现在小盈要去骗煤炭来赚钱。

12. 皮鞋-花瓶*"leather shoe-vase"*

你可以看到一个皮鞋。现在小安要去穿上【皮鞋】出门。

你可以看到一个花瓶。现在小安要去穿上【皮鞋】出门。

你可以看到一个皮鞋。现在小安要去穿上皮鞋出门。

你可以看到一个花瓶。现在小安要去穿上皮鞋出门。

你可以看到一个皮鞋。现在小安要去改善【皮鞋】出门。

你可以看到一个花瓶。现在小安要去改善【皮鞋】出门。

你可以看到一个皮鞋。现在小安要去改善皮鞋出门。

你可以看到一个花瓶。现在小安要去改善皮鞋出门。

13. 老鼠-苹果*"rat-apple"*

你可以看到一个老鼠。现在丹丹要去打死了【老鼠】才能放心。

你可以看到一个苹果。现在丹丹要去打死了【老鼠】才能放心。

你可以看到一个老鼠。现在丹丹要去打死了老鼠才能放心。

你可以看到一个苹果。现在丹丹要去打死了老鼠才能放心。

你可以看到一个老鼠。现在丹丹要去改了【老鼠】才能放心。

你可以看到一个苹果。现在丹丹要去改了【老鼠】才能放心。

你可以看到一个老鼠。现在丹丹要去改了老鼠才能放心。

你可以看到一个苹果。现在丹丹要去改了老鼠才能放心。

14. 口琴-凉鞋*"harmonica-sandal"*

你可以看到一个口琴。现在小宁要去吹【口琴】来练习音乐课的作业。

你可以看到一个凉鞋。现在小宁要去吹【口琴】来练习音乐课的作业。

你可以看到一个口琴。现在小宁要去吹口琴来练习音乐课的作业。

你可以看到一个凉鞋。现在小宁要去吹口琴来练习音乐课的作业。

你可以看到一个口琴。现在小宁要去跑【口琴】来练习音乐课的作业。

你可以看到一个凉鞋。现在小宁要去跑【口琴】来练习音乐课的作业。

你可以看到一个口琴。现在小宁要去跑口琴来练习音乐课的作业。

你可以看到一个凉鞋。现在小宁要去跑口琴来练习音乐课的作业。

15. 瓷器-铅笔*"porcelain-pencil"*

你可以看到一个瓷器。现在莎莎要去洗了【瓷器】再去上班。

你可以看到一个铅笔。现在莎莎要去洗了【瓷器】再去上班。

你可以看到一个瓷器。现在莎莎要去洗了瓷器再去上班。

你可以看到一个铅笔。现在莎莎要去洗了瓷器再去上班。

你可以看到一个瓷器。现在莎莎要去说服【瓷器】再去上班。

你可以看到一个铅笔。现在莎莎要去说服【瓷器】再去上班。

你可以看到一个瓷器。现在莎莎要去说服瓷器再去上班。

你可以看到一个铅笔。现在莎莎要去说服瓷器再去上班。

16. 书包-灯泡*"bag-light bulb"*

你可以看到一个书包。现在欣欣要去整理【书包】再去上学。

你可以看到一个灯泡。现在欣欣要去整理【书包】再去上学。

你可以看到一个书包。现在欣欣要去整理书包再去上学。

你可以看到一个灯泡。现在欣欣要去整理书包再去上学。

你可以看到一个书包。现在欣欣要去规定【书包】再去上学。

你可以看到一个灯泡。现在欣欣要去规定【书包】再去上学。

你可以看到一个书包。现在欣欣要去规定书包再去上学。

你可以看到一个灯泡。现在欣欣要去规定书包再去上学。

17. 地毯-树苗*"carpet-sapling"*

你可以看到一个地毯。现在小怡要去打扫【地毯】再去上班。

你可以看到一些树苗。现在小怡要去打扫【地毯】再去上班。

你可以看到一个地毯。现在小怡要去打扫地毯再去上班。

你可以看到一些树苗。现在小怡要去打扫地毯再去上班。

你可以看到一个地毯。现在小怡要去运行【地毯】再去上班。

你可以看到一些树苗。现在小怡要去运行【地毯】再去上班。

你可以看到一个地毯。现在小怡要去运行地毯再去上班。

你可以看到一些树苗。现在小怡要去运行地毯再去上班。

18. 石头-书法*"stone-calligraphy"*

你可以看到一个石头。现在小杨要去捡了【石头】回家玩。

你可以看到一个书法。现在小杨要去捡了【石头】回家玩。

你可以看到一个石头。现在小杨要去捡了石头回家玩。

你可以看到一个书法。现在小杨要去捡了石头回家玩。

你可以看到一个石头。现在小杨要去负责【石头】回家玩。

你可以看到一个书法。现在小杨要去负责【石头】回家玩。

你可以看到一个石头。现在小杨要去负责石头回家玩。

你可以看到一个书法。现在小杨要去负责石头回家玩。

19. 衣服-电视*"cloth-television"*

你可以看到一个衣服。现在小军要去问候【衣服】再去看病。

你可以看到一个电视。现在小军要去问候【衣服】再去看病。

你可以看到一个衣服。现在小军要去问候衣服再去看病。

你可以看到一个电视。现在小军要去问候衣服再去看病。

你可以看到一个衣服。现在小军要去限制【衣服】再去看病。

你可以看到一个电视。现在小军要去限制【衣服】再去看病。

你可以看到一个衣服。现在小军要去限制衣服再去看病。

你可以看到一个电视。现在小军要去限制衣服再去看病。

20. 空姐-篮球*"air stewardess-basketball"*

你可以看到一个空姐。现在兰萍要去问候【空姐】再去找自己的座位。

你可以看到一个篮球。现在兰萍要去问候【空姐】再去找自己的座位。

你可以看到一个空姐。现在兰萍要去问候空姐再去找自己的座位。

你可以看到一个篮球。现在兰萍要去问候空姐再去找自己的座位。

你可以看到一个空姐。现在兰萍要去承认【空姐】再去找自己的座位。

你可以看到一个篮球。现在兰萍要去承认【空姐】再去找自己的座位。

你可以看到一个空姐。现在兰萍要去承认空姐再去找自己的座位。

你可以看到一个篮球。现在兰萍要去承认空姐再去找自己的座位。

21. 篮子-空调*"basket-air conditioner"*

你可以看到一个篮子。现在淑仪要去打开【篮子】拿出东西来吃。

你可以看到一个空调。现在淑仪要去打开【篮子】拿出东西来吃。

你可以看到一个篮子。现在淑仪要去打开篮子拿出东西来吃。

你可以看到一个空调。现在淑仪要去打开篮子拿出东西来吃。

你可以看到一个篮子。现在淑仪要去维护【篮子】拿出东西来吃。

你可以看到一个空调。现在淑仪要去维护【篮子】拿出东西来吃。

你可以看到一个篮子。现在淑仪要去维护篮子拿出东西来吃。

你可以看到一个空调。现在淑仪要去维护篮子拿出东西来吃。

22. 电器-医院*"electrical appliances-hospital"*

你可以看到一个电器。现在老张要去维修【电器】然后再去睡觉。

你可以看到一个医院。现在老张要去维修【电器】然后再去睡觉。

你可以看到一个电器。现在老张要去维修电器然后再去睡觉。

你可以看到一个医院。现在老张要去维修电器然后再去睡觉。

你可以看到一个电器。现在老张要去邀请【电器】然后再去睡觉。

你可以看到一个医院。现在老张要去邀请【电器】然后再去睡觉。

你可以看到一个电器。现在老张要去邀请电器然后再去睡觉。

你可以看到一个医院。现在老张要去邀请电器然后再去睡觉。

23. 书橱-食物*"shelving-food"*

你可以看到一个书橱。现在嘉玲要去打扫【书橱】再去上班。

你可以看到一个食物。现在嘉玲要去打扫【书橱】再去上班。

你可以看到一个书橱。现在嘉玲要去打扫书橱再去上班。

你可以看到一个食物。现在嘉玲要去打扫书橱再去上班。

你可以看到一个书橱。现在嘉玲要去表扬【书橱】再去上班。

你可以看到一个食物。现在嘉玲要去表扬【书橱】再去上班。

你可以看到一个书橱。现在嘉玲要去表扬书橱再去上班。

你可以看到一个食物。现在嘉玲要去表扬书橱再去上班。

24. 树懒-地铁*"sloth-metro"*

你可以看到一个树懒。现在沛雯要去饲养【树懒】因为他是动物饲养员。

你可以看到一个地铁。现在沛雯要去饲养【树懒】因为他是动物饲养员。

你可以看到一个树懒。现在沛雯要去饲养树懒因为他是动物饲养员。

你可以看到一个地铁。现在沛雯要去饲养树懒因为他是动物饲养员。

你可以看到一个树懒。现在沛雯要去模拟【树懒】因为他是动物饲养员。

你可以看到一个地铁。现在沛雯要去模拟【树懒】因为他是动物饲养员。

你可以看到一个树懒。现在沛雯要去模拟树懒因为他是动物饲养员。

你可以看到一个地铁。现在沛雯要去模拟树懒因为他是动物饲养员。

25. 灯塔-书本*"light house-book"*

你可以看到一个灯塔。现在佳婉要去控制【灯塔】因为他是工作人员。

你可以看到一个书本。现在佳婉要去控制【灯塔】因为他是工作人员。

你可以看到一个灯塔。现在佳婉要去控制灯塔因为他是工作人员。

你可以看到一个书本。现在佳婉要去控制灯塔因为他是工作人员。

你可以看到一个灯塔。现在佳婉要去锻炼【灯塔】因为他是工作人员。

你可以看到一个书本。现在佳婉要去锻炼【灯塔】因为他是工作人员。

你可以看到一个灯塔。现在佳婉要去锻炼灯塔因为他是工作人员。

你可以看到一个书本。现在佳婉要去锻炼灯塔因为他是工作人员。

26. 铅球-词典*"shot put-dictionary"*

你可以看到一个铅球。现在小袁要扔【铅球】来锻炼身体。

你可以看到一个词典。现在小袁要扔【铅球】来锻炼身体。

你可以看到一个铅球。现在小袁要扔铅球来锻炼身体。

你可以看到一个词典。现在小袁要扔铅球来锻炼身体。

你可以看到一个铅球。现在小袁要维修【铅球】来锻炼身体。

你可以看到一个词典。现在小袁要维修【铅球】来锻炼身体。

你可以看到一个铅球。现在小袁要维修铅球来锻炼身体。

你可以看到一个词典。现在小袁要维修铅球来锻炼身体。

27. 凉席-口红*“mat-lipstick”*

你可以看到一个凉席。现在翠萍要去铺上【凉席】然后再睡觉。

你可以看到一个口红。现在翠萍要去铺上【凉席】然后再睡觉。

你可以看到一个凉席。现在翠萍要去铺上凉席然后再睡觉。

你可以看到一个口红。现在翠萍要去铺上凉席然后再睡觉。

你可以看到一个凉席。现在翠萍要去蒸上【凉席】然后再睡觉。

你可以看到一个口红。现在翠萍要去蒸上【凉席】然后再睡觉。

你可以看到一个凉席。现在翠萍要去蒸上凉席然后再睡觉。

你可以看到一个口红。现在翠萍要去蒸上凉席然后再睡觉。

28. 卡通-啤酒*“cartoon-beer”*

你可以看到一个卡通。现在燕妮要去看【卡通】然后再去做作业。

你可以看到一个啤酒。现在燕妮要去看【卡通】然后再去做作业。

你可以看到一个卡通。现在燕妮要去看卡通然后再去做作业。

你可以看到一个啤酒。现在燕妮要去看卡通然后再去做作业。

你可以看到一个卡通。现在燕妮要去改变【卡通】然后再去做作业。

你可以看到一个啤酒。现在燕妮要去改变【卡通】然后再去做作业。

你可以看到一个卡通。现在燕妮要去改变卡通然后再去做作业。

你可以看到一个啤酒。现在燕妮要去改变卡通然后再去做作业。

29. 花生-老师*"peanut-teacher"*

你可以看到一个花生。现在丽虹要去剥了【花生】再去做饭。

你可以看到一个老师。现在丽虹要去剥了【花生】在院子里。

你可以看到一个花生。现在丽虹要去剥了花生在院子里。

你可以看到一个啤酒。现在丽虹要去剥了花生在院子里。

你可以看到一个老师。现在丽虹要关掉【花生】在院子里。

你可以看到一个啤酒。现在丽虹要关掉【花生】在院子里。

你可以看到一个老师。现在丽虹要关掉花生在院子里。

你可以看到一个啤酒。现在丽虹要关掉花生在院子里。

30. 小猫-玫瑰*"kitten-rose"*

你可以看到一个小猫。现在立平要去收养【小猫】来养在家里。

你可以看到一个玫瑰。现在立平要去收养【小猫】来养在家里。

你可以看到一个小猫。现在立平要去收养小猫来养在家里。

你可以看到一个玫瑰。现在立平要去收养小猫来养在家里。

你可以看到一个小猫。现在立平要去品尝【小猫】来养在家里。

你可以看到一个玫瑰。现在立平要去品尝【小猫】来养在家里。

你可以看到一个小猫。现在立平要去品尝小猫来养在家里。

你可以看到一个玫瑰。现在立平要去品尝小猫来养在家里。

31. 鲍鱼-报纸*"abalone-newspaper"*

你可以看到一个鲍鱼。现在思雨要吃了【鲍鱼】然后再去海边玩。

你可以看到一个报纸。现在思雨要吃了【鲍鱼】然后再去海边玩。

你可以看到一个鲍鱼。现在思雨要吃了鲍鱼然后再去海边玩。

你可以看到一个报纸。现在思雨要吃了鲍鱼然后再去海边玩。

你可以看到一个鲍鱼。现在思雨要参考【鲍鱼】然后再去海边玩。

你可以看到一个报纸。现在思雨要参考【鲍鱼】然后再去海边玩。

你可以看到一个鲍鱼。现在思雨要参考鲍鱼然后再去海边玩。

你可以看到一个报纸。现在思雨要参考鲍鱼然后再去海边玩。

32. 手袋-手机*"handbag-phone"*

你可以看到一个手袋。现在小真要去提上【手袋】去看演出。

你可以看到一个手机。现在小真要去提上【手袋】去看演出。

你可以看到一个手袋。现在小真要去提上手袋去看演出。

你可以看到一个手机。现在小真要去提上手袋去看演出。

你可以看到一个手袋。现在小真要去成为【手袋】去看演出。

你可以看到一个手机。现在小真要去成为【手袋】去看演出。

你可以看到一个手袋。现在小真要去成为手袋去看演出。

你可以看到一个手机。现在小真要去成为手袋去看演出。

33. 面包-面条*"bread-noodle"*

你可以看到一个面包。现在小森要去烤上【面包】再去做饭。

你可以看到一个面条。现在小森要去烤上【面包】再去做饭。

你可以看到一个面包。现在小森要去烤上面包再去做饭。

你可以看到一个面条。现在小森要去烤上面包再去做饭。

你可以看到一个面包。现在小森要去参加【面包】再去做饭。

你可以看到一个面条。现在小森要去参加【面包】再去做饭。

你可以看到一个面包。现在小森要去参加面包再去做饭。

你可以看到一个面条。现在小森要去参加面包再去做饭。

34. 牛肉-牛奶*"beef-milk"*

你可以看到一个牛肉。现在小娜要去水煮【牛肉】当作午餐。

你可以看到一个牛奶。现在小娜要去水煮【牛肉】当作午餐。

你可以看到一个牛肉。现在小娜要去水煮牛肉当作午餐。

你可以看到一个牛奶。现在小娜要去水煮牛肉当作午餐。

你可以看到一个牛肉。现在小娜要去租【牛肉】当作午餐。

你可以看到一个牛奶。现在小娜要去租【牛肉】当作午餐。

你可以看到一个牛肉。现在小娜要去租牛肉当作午餐。

你可以看到一个牛奶。现在小娜要去租牛肉当作午餐。

35. 蜂蜜-风扇*"honey-fan"*

你可以看到一个蜂蜜。现在小佩要去喝了【蜂蜜】然后再去吃晚餐。

你可以看到一个风扇。现在小佩要去喝了【蜂蜜】然后再去吃晚餐。

你可以看到一个蜂蜜。现在小佩要去喝了蜂蜜然后再去吃晚餐。

你可以看到一个风扇。现在小佩要去喝了蜂蜜然后再去吃晚餐。

你可以看到一个蜂蜜。现在小佩要去打印【蜂蜜】然后再去吃晚餐。

你可以看到一个风扇。现在小佩要去打印【蜂蜜】然后再去吃晚餐。

你可以看到一个蜂蜜。现在小佩要去打印蜂蜜然后再去吃晚餐。

你可以看到一个风扇。现在小佩要去打印蜂蜜然后再去吃晚餐。

36. 电脑-电线*"computer-wire"*

你可以看到一个电脑。现在小龙要去使用【电脑】来完成作业。

你可以看到一个电线。现在小龙要去使用【电脑】来完成作业。

你可以看到一个电脑。现在小龙要去使用电脑来完成作业。

你可以看到一个电线。现在小龙要去使用电脑来完成作业。

你可以看到一个电脑。现在小龙要去批评【电脑】来完成作业。

你可以看到一个电线。现在小龙要去批评【电脑】来完成作业。

你可以看到一个电脑。现在小龙要去批评电脑来完成作业。

你可以看到一个电线。现在小龙要去批评电脑来完成作业。

37. 酒杯-酒精*"wine glass-surgical spirit"*

你可以看到一个酒杯。现在小希要去清洗【酒杯】再去喝酒。

你可以看到一个酒精。现在小希要去清洗【酒杯】再去喝酒。

你可以看到一个酒杯。现在小希要去清洗酒杯再去喝酒。

你可以看到一个酒精。现在小希要去清洗酒杯再去喝酒。

你可以看到一个酒杯。现在小希要去攀爬【酒杯】再去喝酒。

你可以看到一个酒精。现在小希要去攀爬【酒杯】再去喝酒。

你可以看到一个酒杯。现在小希要去攀爬酒杯再去喝酒。

你可以看到一个酒精。现在小希要去攀爬酒杯再去喝酒。

38. 冰块-冰箱*"ice cube-refrigerator"*

你可以看到一个冰块。现在小诗要去制作【冰块】来获得水。

你可以看到一个冰箱。现在小诗要去制作【冰块】来获得水。

你可以看到一个冰块。现在小诗要去制作冰块来获得水。

你可以看到一个冰箱。现在小诗要去制作冰块来获得水。

你可以看到一个冰块。现在小诗要去责备【冰块】来获得水。

你可以看到一个冰箱。现在小诗要去责备【冰块】来获得水。

你可以看到一个冰块。现在小诗要去责备冰块来获得水。

你可以看到一个冰箱。现在小诗要去责备冰块来获得水。

39.雨衣-雨伞*"raincoat-umbrella"*

你可以看到一个雨衣。现在舒雅要穿上【雨衣】才能出门因为外面下起了大雨。

你可以看到一个雨伞。现在舒雅要穿上【雨衣】才能出门因为外面下起了大雨。

你可以看到一个雨衣。现在舒雅要穿上雨衣才能出门因为外面下起了大雨。

你可以看到一个雨伞。现在舒雅要穿上雨衣才能出门因为外面下起了大雨。

你可以看到一个雨衣。现在舒雅要问候【雨衣】才能出门因为外面下起了大雨。

你可以看到一个雨伞。现在舒雅要问候【雨衣】才能出门因为外面下起了大雨。

你可以看到一个雨衣。现在舒雅要问候雨衣才能出门因为外面下起了大雨。

你可以看到一个雨伞。现在舒雅要问候雨衣才能出门因为外面下起了大雨。

40.乌梅-乌龟*"dark plum-tortoise"*

你可以看到一个乌梅。现在宁宁要去吃了【乌梅】再去看电影。

你可以看到一个乌龟。现在宁宁要去吃了【乌梅】再去看电影。

你可以看到一个乌梅。现在宁宁要去吃了乌梅再去看电影。

你可以看到一个乌龟。现在宁宁要去吃了乌梅再去看电影。

你可以看到一个乌梅。现在宁宁要去打了【乌梅】再去看电影。

你可以看到一个乌龟。现在宁宁要去打了【乌梅】再去看电影。

你可以看到一个乌梅。现在宁宁要去打了乌梅再去看电影。

你可以看到一个乌龟。现在宁宁要去打了乌梅再去看电影。